FACULTAD DE ESTUDIOS ESTADÍSTICOS

MÁSTER EN MINERÍA DE DATOS E INTELIGENCIA DE NEGOCIOS

Curso 2018/2019

Trabajo de Fin de Máster

TITULO: EVALUACIÓN DE ALGORITMOS PARA LA DETECCIÓN DE IMPAGO DE LENDINGCLUB

***Alumno*: Marta Palau Payeras**

***Tutor*: Lorenzo Escot Mangas**

Julio de 2019

UNIVERSIDAD COMPLUTENSE
MADRID

Índice

Índice de tablas y gráficos

1. Introducción del trabajo

El papel que adoptó el sector financiero en la crisis de la economía mundial que estalló con la caída del Lehman Brothers en septiembre del 2008 dio lugar al posicionamiento del sector en el punto de mira. Debido a la carencia que hubo entre la regulación y la supervisión financiera, al cúmulo de ineficiencias en sistemas de incentivos, a las prácticas bancarias deficientes y a la mala gestión del riesgo, el sector financiero fue de los primeros en ser intervenidos para combatir la recesión. Se impusieron políticas fuertemente restrictivas que llevaron al racionamiento del crédito, incrementando la exigencia de condiciones ante nuevas peticiones lo cual supuso una mayor restricción de financiación a muchas empresas y particulares (Bruno, 2017).

Además, cabe destacar el comportamiento de los tipos de interés y las rentabilidades desde entonces, las cuales se sitúan cercanas al 0% para inversiones en fondos conservadores llevando incluso a incurrir en pérdidas (De La Cruz, 2018).

Esta breve explicación permite encarrilar las inquietudes que orientan este trabajo a realizar un análisis del funcionamiento de una empresa que surge como alternativa para dar opción de financiación a aquellos particulares y pequeñas empresas que han sido excluidos del mercado de crédito tradicional. En el 2007, surge la empresa LendingClub en Estados Unidos para darle un giro al sector bancario y conseguir hacer que el crédito sea más accesible y que invertir sea más fructífero.

Se trata de una empresa que sigue la línea de economía colaborativa, las cuales van asociadas al concepto de *startup* digital que realmente hacen referencia al *peer-to-peer* de siempre pero que han dado el salto a la nueva era tecnológica y tienen como principal vía los canales digitales (desde funcionamiento online de página web hasta las novedosas *apps*). Se trata de un modelo de financiación basado en la intermediación entre aquellos que necesitan crédito y aquellos que buscan invertir su dinero a cambio de rentabilidad, utilizando como medio de comunicación entre las partes una plataforma digital. Además, el uso de este canal permite operar con bajos costes y evitar precios abusivos, centrándose así en dar un buen servicio ajustado a las necesidades del consumidor. Su principal función es dar acceso a inversores a peticiones de crédito de usuarios (préstamos, auto refinanciación, préstamos de empresas etc.), permitiéndoles obtener una rentabilidad que varía según sea el riesgo asociado al portfolio en el que se deposita el dinero.

1.1 Funcionamiento de LendingClub

La empresa se define como líder mundial de mercado online cuyo objetivo es poner en contacto inversores y prestatarios en un entorno de transformación de la industria bancaria para hacerla más transparente, eficiente y cercana al cliente. A continuación damos una pincelada de los aspectos más importantes del funcionamiento de la empresa y de las etapas de los créditos y las inversiones que gestionan:

- Los clientes interesados en adquirir un préstamo realizan una solicitud online, facilitando los detalles y la información personal requerida por la empresa.
- Se evalúa la solicitud y la información personal del solicitante de forma que se caracteriza la inversión, adjudicándole su riesgo y su tipo de interés y se presenta la oferta a inversores cualificados.

- Los inversores, desde particulares hasta instituciones, seleccionan préstamos en los que les interesa invertir y generar rendimientos. Sus decisiones de diversificación dependen de su tolerancia al riesgo, sus objetivos de composición de cartera de riesgo y el horizonte temporal que estén dispuestos a asumir.

1.1.1 Cómo gana dinero LendingClub

LendingClub, como empresa intermediaria obtiene beneficios de las tasas que pagan ambos, prestatario e inversor. Las tasas de los inversores se obtienen cuando estos reciben pagos, por lo que su beneficio está estrictamente condicionado a su flujo de caja. Dicho esto, los beneficios se obtienen por las siguientes vías:

- Ingresos por las tasas de creación de crédito personal mediante las cuales LendingClub carga entre el 1 y el 6% de la cantidad total solicitada. La tasa varía según sea el rating del individuo y la información revelada por su aplicación.
- Ingresos procedentes del importe total recibido por beneficios de los inversores del 1% por cada pago que se les efectúa. Si los prestatarios dejan de hacer un pago, los inversores no abonarán este importe a LendingClub.
- Si el prestatario no paga o se retrasa en sus pagos, LendingClub utiliza las mismas medidas que el sector bancario clásico para tratar de recuperarlos. Cuando tengan que mediar para que los inversores reciban la cuota, si finalmente sus actuaciones generan resultados satisfactorios se cargará un importe en torno el 30-40% de la cantidad que se consiga recuperar (dependiendo de si requieren intervención legal o no). En el caso que no se recolecte ningún pago, la empresa no carga tasa de recobro.

Por ello, es importante que se realicen filtros de calidad a todas las aplicaciones realizadas para adquirir un préstamo ya que es necesario dentro de lo posible asegurar retorno a los inversores y así que estos puedan reinvertir ese dinero, haciendo crecer el negocio.

1.1.2 Rating de LendingClub

Para evaluar el riesgo y determinar los tipos de interés a los préstamos aprobados se realizan modelos que incluyen como input todas las características de los prestatarios que puedan ser determinantes para asociar probabilidades de incumplimiento. Dependiendo de la categorización que reciben los prestatarios se determinará la facilidad con la que podrán adquirir un préstamo, una nueva tarjeta de crédito o refinanciación de deuda. Se asignará una puntuación que determinará la probabilidad de impago de la persona.

Los factores más determinantes, ordenados de mayor a menor importancia según presenta LendingClub en su página web son: historial de pagos, antigüedad y tipo de créditos que solicita, porcentaje de uso dado el límite total, balance total o balance de deuda, comportamiento reciente de crédito, consultas realizadas al individuo y finalmente el crédito disponible. Toda esta información es principalmente adquirida de reportes oficiales de crédito que deben ser facilitados en el momento de solicitud.

La empresa, además, asegura que factores como raza, color, religión, nacionalidad, sexo, estado civil, edad, salario, ocupación, formación/títulos, historial de empleo, lugar de residencia o incluso el total de bienes inmuebles bajo disposición no se toman en consideración para evaluar los perfiles de los solicitantes. De esta forma aseguran utilizar modelos consistentes y elaborados, tratando de asegurar cierto retorno y minimizando el riesgo asumido.

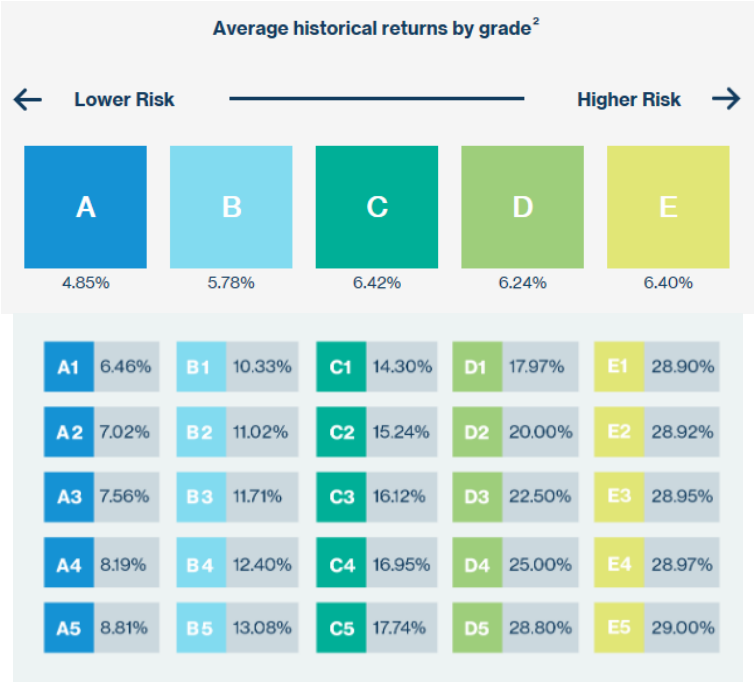
Es importante mencionar que los inversores no están invirtiendo su dinero directamente en un préstamo de un individuo, sino que realmente se invierte en lo que ellos llaman *notes*. Cada préstamo aprobado es dividido por LendingClub en pequeñas fracciones, dando así la posibilidad a los inversores que inviertan en las *notes* que les interesen dependiendo de su perfil y su objetivo.

De esta forma se lleva a cabo una clasificación previa en categorías de la A hasta la E, reflejando el riesgo que contiene cada una y determinando también el tipo de interés que generarán. Este tipo de interés es fruto de dos partes, una fija y la otra variable que depende del riesgo y de las condiciones de mercado:

Tipo de interés= LC tipo interés base + ajuste por riesgo y volatilidad

La parte variable permite cubrir las pérdidas esperadas y ajustar mejor el retorno a mayor riesgo afrontado, por lo que dentro de cada una de las categorías de la A a la E se crean 5 subcategorías más.

Figura 1: Relación entre Riesgo y Rentabilidad reflejada con la Puntuación A-E



LOAN GRADE	SUB-GRADE	LENDING CLUB BASE RATE	ADJUSTMENT FOR RISK & VOLATILITY	INTEREST RATE
F	1	5.05%	24.30%	29.35%
	2	5.05%	24.64%	29.69%
	3	5.05%	25.12%	30.17%
	4	5.05%	25.60%	30.65%
	5	5.05%	25.70%	30.75%
G	1	5.05%	25.74%	30.79%
	2	5.05%	25.79%	30.84%
	3	5.05%	25.84%	30.89%
	4	5.05%	25.89%	30.94%
	5	5.05%	25.94%	30.99%

Fuente: LendingClub

Al final de la figura 1 puede apreciarse como, independientemente de la categoría, la parte fija del tipo de interés es la misma para todos y lo que varía es la parte de ajuste por volatilidad y riesgo.

Como se puede observar, tenemos cinco grupos de A hasta E, con cinco subgrupos más en cada uno. Además, existen los grupos F y G con tal nivel de riesgo asociado que únicamente están disponibles para los inversores que tienen como objetivo la especulación debido a que la probabilidad de perder todo el dinero invertido en ellos es muy elevada.

Cuando una aplicación pasa los primeros filtros y se da por válida, las características de la aplicación se analizan más en profundidad con el modelo de puntuación propio de la empresa que lleva a rechazar o aprobar la propuesta. El modelo inicial da una puntuación a la aplicación, la cual combinada con otras características de la solicitud dan lugar a la segunda fase de puntuación donde se utiliza el modelo creado por la empresa. Este modelo está compuesto por un algoritmo que analiza el comportamiento del prestatario, su puntuación inicial y otras características adjuntas en la solicitud; de esta forma se lleva a cabo un ranking del 1 al 25 para las categorías y subcategorías previamente presentadas (de la A.1 a la E.5). Con esto se atribuyen distintas puntuaciones definitivas a las *notes* disponibles para que los inversores depositen su dinero en ellas según sea su criterio en cuanto a tolerancia de riesgo y sus objetivos de inversión.

Teniendo en cuenta todo esto, los inversores deben decidir cómo componer su porfolio de inversión, por lo que tienen dos formas de hacerlo: por una parte pueden elegir bajo criterio propio en qué *notes* invertir y por otra parte pueden automatizar este procedimiento estableciendo requisitos de rentabilidad y riesgo a partir de los cuales la empresa propondrá recomendaciones de composición óptima de cartera (la última palabra la tiene siempre el inversor).

1.1.3 Caracterización de las solicitudes recibidas por LendingClub

Si pasamos a valorar las solicitudes que recibe LendingClub, es importante fijarnos en los requisitos para poder solicitar un préstamo:

- Ciudadanos de Estados Unidos, residentes permanentes o con visados de larga duración.
- Edad mínima de 18 años.
- Propietario de una cuenta bancaria verificable.

Además, para aceptar la solicitud realizada es necesario que se cumpla también:

- Puntuación en el modelo de inicial mínima fijada de 620 puntos.
- Disposición mínima de historial crediticio de 3 años.
- Un máximo de ratio deuda/ingresos del 40%.

En ocasiones, cuando una solicitud no alcanza la puntuación mínima LendingClub permite añadir un co-prestatario el cual puede ayudar en la calificación obtenida. Simplemente se deberá indicar en la solicitud y añadir la información requerida sobre el mismo.

El importe que puede solicitarse está entre $1.000 y $40.000, con unos intereses que varían entre el 6,95% hasta el 35,89% como ya hemos visto. La empresa carga a los individuos por el tipo de interés anual del crédito que pagan a los inversores y por la tasa de creación de cuenta de crédito (entre el 1 y el 6% según la puntuación crediticia del solicitante). Todos los créditos personales tienen tasas y pagos mensuales fijos. Los plazos

de solicitud son de 30 o 60 meses, con disponibilidad de efectuar la devolución de la totalidad del dinero adquirido más los intereses en menos tiempo. No se penalizará a los prestatarios por ello y los inversores reciben su parte menos el 1% de tasa de servicio cargada como si fuera el mismo porcentaje mensual estipulado. En cuanto a los pagos fuera de plazo, se concede un margen de 15 días a partir del día en que se debería llevar a cabo el pago correspondiente. Una vez superado este periodo se penaliza con tasas y otros recargos para compensar a los inversores que están respaldando el crédito.

1.2 Justificación del trabajo

Si analizamos más detenidamente la funcionalidad de LendingClub vemos que se trata de una plataforma online en la que aquellos individuos que han sido descartados por el sector financiero tradicional acuden para solicitar financiación. Por ello, es necesario generar rentabilidades más altas que atraigan a los inversores y compensen el mayor riesgo asumido al depositar el dinero en sus productos.

La empresa utiliza su propio modelo de puntuación, el cual cabe ver si es realmente objetivo y persigue que los inversores tengan una visión real del riesgo en el que se adentran al realizar ciertas inversiones. Nosotros llevaremos a cabo diferentes pruebas usando diferentes algoritmos para ver si conseguimos mejorar la detección de préstamos impagados. Modelizaremos con la clásica regresión logística y con los algoritmos de *Machine Learning* para evaluar que alternativa se adapta mejor a los datos y permite llevar a cabo un buen filtro inicial de solicitudes. Finalmente, para el mejor modelo, evaluaremos el contenido y analizaremos cuáles son las características individuales más importantes para determinar si es favorable o no conceder un préstamo a cada uno de los solicitantes. Además, trataremos de verificar si realmente procede utilizar toda la información individual facilitada en el momento de solicitud del préstamo y si no se utilizan otros criterios para determinar la decisión de concesión.

1.3 Estado del arte del estudio

Hay diferentes autores que han realizado trabajos que pueden servirnos como inspiración o como base de la cual partir. Por ejemplo, en (Haotian, Ziyuan, Tianyu y Zhou 2015) utilizan técnicas de minería de datos para obtener predicciones sobre la probabilidad de impago de crédito. Existen numerosos estudios en los que se plantea la modelización de este tipo de comportamientos de impago, fraude, etc. buscando detectar riesgos y prevenir así situaciones desfavorables como la que observamos en la última crisis financiera, especialmente si nos fijamos en el sector bancario. Por otra parte, en Valle (2015) se utilizan otro tipo de modelos para obtener un objetivo similar pero más enfocado al sector financiero tradicional. El informe aportado por (Skantzoa y Catelein 2016) es específico y muy útil para adentrarse en el modelo de probabilidad, ya que proporciona pautas para asegurar la calidad y el desarrollo adecuado del análisis de riesgos. También se utiliza el trabajo de final de Máster realizado por (Company, 2018) que genera un modelo de clasificación para una empresa de reinversión de préstamo usando distintas técnicas de modelización. De aquí se cogen algunas ideas de lo que se suele extraer de los análisis aplicados a este tipo de empresas que abastecen el mercado de crédito secundario. En cuanto a la metodología de desarrollo técnico de modelos utilizamos el manual de Saddiqi (2005).

Todos estos estudios especificados y referenciados han servido como guía para comprender el procedimiento que normalmente se sigue al realizar este tipo de análisis. Nosotros establecemos una base similar de trabajo pero iremos construyendo nuestra propia estructura concorde a los objetivos que a continuación fijamos. Además, casi todas las técnicas utilizadas en el trabajo han sido aprendidas en el Máster de Minería de Datos e Inteligencia de Negocios, por lo que también se utiliza material adquirido durante dicha formación.

1.4 Objetivos del trabajo

Los objetivos principales de este proyecto son los siguientes:

- **Contrastaremos si la información individual de las solicitudes facilitada por la empresa permite componer un buen modelo para seleccionar aquellos individuos a los que sí conviene conceder un préstamo.**

LendingClub es una empresa en la que se obtienen beneficios con origen en las tasas y comisiones que cobra a los prestatarios y a los inversores, además de otros beneficios extraordinarios generados por situaciones de mora de pagos. Además, como el riesgo de impago es transferido mayoritariamente a los inversores es importante comprobar que es posible establecer con la información facilitada en las solicitudes unos criterios adecuados y que realmente permitan presentar una puntuación transparente y alineada con los intereses de los inversores.

- **Crearemos diferentes modelos utilizando diferentes algoritmos para conseguir determinar cuál es el más apropiado. Evaluaremos si, con las características de las que dispone LendingClub cuando se realizan las solicitudes es posible conseguir realizar un buen filtrado de malos y buenos pagadores.**

Utilizaremos toda la información facilitada por la propia empresa sobre los solicitantes de crédito como si fuéramos a invertir y escoger así aquellos individuos a los que conviene aceptar o rechazar la solicitud de préstamo. Normalmente, se utiliza la regresión logística por su buen ajuste y su facilidad en la interpretación de los parámetros y los efectos en la variable objetivo. Trataremos de utilizar algoritmos más potentes y complejos de *Machine Learning* y evaluaremos si compensa la pérdida de interpretabilidad que suponen exigiendo cierto nivel de mejora en el ajuste y fiabilidad de los modelos con respecto a la regresión.

- **Finalmente evaluaremos para el mejor modelo cuál es su composición y qué variables y comportamiento en cada una de ellas determinan que un individuo vaya a ser buen o mal pagador para entender cómo funcionan los modelos de Credit Scoring.**

Para ello utilizaremos derivadas parciales en las que, una vez se tiene clara la composición del modelo, evaluaremos como varía para cada una de las características de entrada la probabilidad de impago con respecto a un individuo medio. Así podremos ver como varía la probabilidad de impago y a ello asociar que sea conveniente conceder o no el préstamo a una solicitud dadas sus características utilizando una atribución de puntos. Finalmente se suman o se restan los puntos asociados al individuo dados todos los valores que se generen en cada variable de entrada y así se establece un rango a partir del cual se concederá o no el crédito.

2 Metodología y marco teórico del estudio

La realización del trabajo se basará en el esquema de la metodología SEMMA, la cual está compuesta por cinco fases:

1. *Sample*: fase que consiste en seleccionar la muestra representativa de los datos que nos permitirá estudiar y desarrollar el procedimiento en búsqueda de identificar resultados al objetivo planteado. Disponemos de una base de datos grande pero con una proporción pequeña de eventos. Es por ello que aplicaremos la técnica de sobre-muestreo, utilizando así parte de los datos iniciales para modelizar y parte para validar.
2. *Explore*: analizaremos, entenderemos y estudiaremos de forma profunda las variables disponibles. Se trata de la base del funcionamiento del modelo que posteriormente crearemos. Lo primero que tendremos que hacer es identificar y definir correctamente la variable objetivo. Una vez tengamos claro el primer paso básico, nos centraremos en ver aquellas variables input que puedan aportar valor y ser útiles para conseguir nuestros objetivos. Tendremos que tratar los datos y realizar la revisión de los siguientes puntos: modificación de posibles errores, eliminar datos atípicos y anómalos, tratamiento oportuno de datos faltantes etc. Además, realizaremos análisis univariante y bivariante para ver el aporte de las variables y si existiera necesidad de transformarlas para mejorarlas o también valorar la posibilidad de obtención de nuevas variables a partir de las que ya tenemos. Todo esto forma parte de la preparación de los datos para lo que será la modelización, por lo que es necesario dedicar gran parte del tiempo a esta fase del proyecto.
3. *Modify*: a continuación, realizaremos las modificaciones necesarias previamente identificadas para conseguir una mejor relación con la variable objetivo y así obtener un buen modelo (tramificación, reagrupación etc). Evaluaremos con estadísticos apropiados si realmente tienen aporte para el desarrollo del modelo las variables definitivas con las que consideramos oportuno comenzar la etapa de modelización.
4. *Model*: aquí construiremos modelos e identificaremos aquel que permitan plasmar de forma óptima la relación entre la variable objetivo (dependiente) y el resto de las variables input para detectar con la mayor cantidad de aciertos posibles los casos de impago. Por ello, como hemos comentado a priori, utilizaremos las siguientes alternativas de modelización: regresión logística, redes neuronales, *bagging, random forest, gradient boosting, support vector machines* y también realizaremos alguna prueba de ensamblado de modelos.
5. *Asses*: tras obtener diferentes modelos, los evaluamos como de bien o mal predicen, buscando cometer el mínimo error posible.

2.1 Regresión Logística

Se trata una técnica de modelización estadística que estudia la relación entre las variables independientes y la variable objetivo categórica binaria. La técnica modela la probabilidad de que ocurra el evento frente a la de que no ocurra en función de las variables independientes. Se utiliza la función de enlace logística que permite relacionar las variables input con la probabilidad de forma que se mantenga dentro del intervalo (0,1), la cual asume la siguiente relación:

$$P_1 = P(Y = 1|x_1, x_2, \dots x_m) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m)}}$$

$$P_0 = 1 - P_1 = P(Y = 0|x_{1,}\, x_{2,}\, \dots\, x_m) = \frac{e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)}}{1 + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_m x_m)}}$$

$$\log\left(\frac{P_1}{1-P_1}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m$$

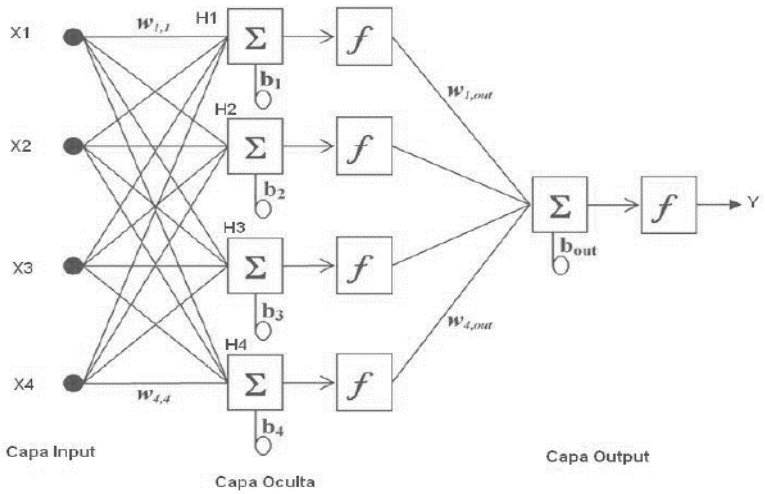
El término $log\ \left(\frac{P_1}{1-P_1}\right)$ es el *logit* y es el logaritmo de la razón de probabilidades u *odds ratio*. Los *odds ratio* son aquellos coeficientes que componen la regresión de los cuales extraemos la utilidad de interpretación de los efectos que tienen las categorías incluidas en el estudio sobre la variable dependiente. Además, cabe comentar que en el caso de la regresión es necesario recurrir a métodos iterativos de optimización ya que no existe una fórmula explícita que permita obtener los parámetros que maximizan la verosimilitud (SAS Technical Support, *s.f.*).

A continuación introducimos los algoritmos de *Machine Learning* que utilizaremos en el trabajo. Todas las explicaciones realizadas están basadas en las explicaciones dadas por (Portela, 2019) por lo que a continuación realizamos un breve resumen para cada técnica:

2.2 Redes Neuronales

La red neuronal es un algoritmo de *Machine Learning* que imita el funcionamiento de las neuronas del cerebro: cada neurona procesa y combina diferentes estímulos de las otras neuronas con las que están interconectadas.

Figura 2: Estructura de la red neuronal



Fuente: Apuntes Redes Neuronales (Portela, 2019)

La red neuronal está compuesta por diferentes nodos conectados entre sí, organizados en grupos llamados capas. En la figura 2 se representa como la capa input, compuesta por aquellas variables que introducimos para modelizar (x_i), se conecta con la capa oculta (H_j) mediante la función de combinación que otorga pesos (w_{ij}), los cuales representan la interacción entre los nodos de las capas y a cada una de estas combinaciones se le asocia un sesgo (b_{ij}). Posteriormente se aplica la función de activación oportuna (f), la cual impone el límite que se debe sobrepasar antes de llevar a cabo la combinación de todas las relaciones. Finalmente, se aplica la activación de la combinación generada de la capa oculta a la capa output, dando lugar finalmente a la Y. Este proceso, cuantas más variables de entrada y más capas ocultas, más complejo se vuelve y más parámetros se generan, por

lo que se debe controlar el grado de complejidad adecuado según sea la composición del set de datos que se trabaje.

Las redes neuronales se entrenan, es decir, hay que buscar el valor de los pesos (w_{ij}) que mejor ajusta la variable objetivo a partir de las variables input. Esto hace que se requieran una gran cantidad de datos para poder "entrenar" a la red, que gane en experiencia y que sea robusta para encontrar la combinación de parámetros que mejor clasificará. Dicho esto, las redes neuronales requieren especial atención en configurar una buena parametrización de los siguientes componentes:

- Número de nodos.
- Función de activación.
- Algoritmo de optimización (si usamos como algoritmo *backpropagation* podremos jugar con los siguientes parámetros de regularización: *momentum* y *learning rate*).
- Número de Iteraciones máximas, realizando pruebas de *early stopping* (criterio de parada anticipada) ya que en ocasiones permite evitar el sobre ajuste.

2.3 Métodos basados en árboles

Estos algoritmos de aprendizaje automático se basan en combinar la salida de varios árboles mediante el promediado de modelos con la finalidad de mejorar la estabilidad y precisión de los algoritmos.

Los árboles, más concretamente de clasificación para variable objetivo binaria, son una herramienta de modelización que representan una segmentación de los datos a partir de una serie de reglas simples, que se van aplicando de forma jerárquica y secuencial. De esta forma se obtienen una serie de segmentos (los nodos) que contienen subconjuntos de la muestra. Una vez obtenida la segmentación óptima, entendiendo así aquella que da lugar a nodos con comportamiento homogéneo respecto a la variable objetivo y heterogéneos entre ellos, se asigna un valor de predicción a los nodos finales, de forma que todas las observaciones pertenecientes a cada uno de estos serán predichas a partir de dicho valor.

Dentro de estos métodos existen estas tres variantes:

2.3.1 Bagging

Este es el primer método que se utilizó, basado en un conjunto de datos de tamaño N del cual se seleccionan N o n>N observaciones con reemplazamiento (o sin) de los datos originales y se aplica un árbol, del cual se obtienen las predicciones para todas las observaciones originales N. Este proceso se repite las m veces que se decida para finalmente promediar las m predicciones obtenidas.

2.3.2 Random Forest

Este método es una modificación de *bagging* que consiste en incorporar la aleatoriedad en las variables utilizadas para segmentar cada nodo del árbol. El proceso es prácticamente igual con la única diferencia de que cada vez que se abre un nodo seleccionaremos p variables de las k originales y de esas p elegidas, se escoge la mejor para llevar a cabo la partición en ese nodo.

2.3.3 Gradient Boosting

Este tercer método se diferencia de las dos versiones anteriores por ir modificando las predicciones en la dirección de decrecimiento dada por el negativo del gradiente de la función de error. Se basa en repetir la construcción de los árboles modificando ligeramente las predicciones iniciales cada vez, intentando ir minimizando los residuos en un parámetro V. Al plantear muchos árboles de forma repetida conseguimos que se ajusten cada vez más las predicciones a los datos y, a su vez, van corrigiéndose los unos a los otros permitiendo mucha adaptabilidad para la construcción de un buen modelo.

Dada la explicación de los tres métodos, como todos se basan en la construcción de muchos árboles, extraemos que debemos prestar atención a los siguientes parámetros:

- Propiedades de los árboles de clasificación:
  - Número de hojas final o profundidad del árbol.
  - Número de divisiones máximas en cada nodo (lo dejaremos en dos por defecto).
  - P-valor: determina las divisiones en cada nodo; más alto, más estricto frente a la realización de subdivisiones y más sencillos son los árboles.
  - El número de observaciones mínimo que debe de haber en una rama/nodo. Si se establece un número más grande permite evitar sobreajuste (menos varianza) a cambio de algo más de sesgo, frente a reducir el número y ajustar mejor a los datos (menor sesgo, algo más de varianza en los modelos).
- Número de iteraciones m a promediar.
- Tamaño de la muestra n vs N para la que realizaremos los árboles.

Para *bagging*, esto sería todo lo que tendremos en cuenta, mientras que para *random forest*, además, exploraremos el número óptimo de variables p a muestrear en cada nodo.

Para el caso de *gradient boosting* debemos tener en cuenta:

- La constante de regularización que refleja en cuanto se modifica el error cada vez. Cuanto más elevada sea, más rápido converge el algoritmo, por lo que debemos encontrar un valor óptimo para evitar ser ni demasiado bruscos ni demasiado lentos.
- Necesidad de *early stopping* (parada anticipada del algoritmo para evitar sobre ajuste).
- *Stochastic gradient boosting*: podríamos probar si hay diferencias sustanciales entre realizar un procedimiento estocástico y no hacerlo. Este se basa en cada iteración, es decir, en la elaboración de cada árbol, utilizar un archivo de entrenamiento diferente. Se trata de otra opción para paliar el sobre ajuste (está opción está disponible en los paquetes de R).

2.4 Support Vector Machines

Este algoritmo se basa en solucionar un problema de separación lineal de clases con métodos algebraicos mediante la búsqueda del hiperplano de separación. Está compuesto por diferentes modalidades que han ido mejorando progresivamente:

- *Maximal margin*: añade el criterio del separador con máximo margen a la hora de dividir las clases por un hiperplano. Se trata minimizar la siguiente función de forma que nos permita generar el máximo margen hallando el vector de parámetros adecuado:

$$\arg\min_{(\mathbf{w},b)} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{subject to (for any } i = 1, \ldots, n)$$
$$y_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1.$$

- *Soft margin*: se mejora el concepto tratando de dar libertad a la restricción, asumir que la separación perfecta no suele existir y que es necesario permitir que algunas observaciones queden mal clasificadas por los separadores para no incurrir en sobreajuste. Esto se consigue añadiendo un parámetro de residuo y una constante de regularización C de margen, la cual marca el permiso de fallo. La función se ve modificada de la siguiente forma:

$$\arg\min_{\mathbf{w},\xi,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \right\}$$
$$\text{subject to (for any } i = 1, \ldots n)$$
$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

- *Kernel*: última idea introducida permite tratar la separación no lineal entre clases. Este nos permite trabajar con un algoritmo de separación lineal con un problema no lineal trasladándonos a un espacio de dimensión superior donde podamos aplicar la separación lineal. Conseguimos este objetivo introduciendo nuevas funciones de las variables que sean no lineales, aumentando así la dimensión del vector de variables independientes, siendo así más fácil para el algoritmo encontrar se unas funciones estándar a las cuales se asocian los siguientes parámetros:
  - Lineal: contamos exclusivamente con el parámetro C, inverso al permiso de fallo:
$$K(\mathbf{x}_i \cdot \mathbf{x}_j) = \mathbf{x}_i^T \cdot \mathbf{x}_j$$
  - Polinomial: además de C, debemos tener en cuenta el grado del polinomio que usamos para cambiar la dimensión y el parámetro *scale* que marca la escala del polinomio del *kernel*, sirve para normalizar sin tener que modificar los datos:
$$K(\mathbf{x}_i \cdot \mathbf{x}_j) = \left(\gamma \mathbf{x}_i^T \cdot \mathbf{x}_j + r\right)^d$$
  - Función Radial (RBF): en este caso debemos parametrizar C y sigma, que representa la inversa del radio de influencia de las muestras seleccionadas por el modelo como *support vectors*. Lo que hace es definir como de lejos o cerca alcanza la influencia:
$$K(\mathbf{x}_i \cdot \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$
$$= \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \gamma = \frac{1}{2\sigma^2}$$

2.5 Técnicas de Ensamblado

Esta técnica se basa en la combinación de distintos modelos, es decir, combinar los resultados de las predicciones obtenidas con distintos algoritmos y tratar de mejorar así las predicciones. El objetivo es alcanzar robustez por el hecho de que unos modelos corrigen a otros, lo cual además suele ir ligado a una reducción de la varianza en general y difícilmente si los modelos son buenos empeoren con el ensamblado.

2.6 Concepto Validación Cruzada Repetida

Para evaluar la bondad de ajuste de los modelos se utiliza esta técnica que reduce la dependencia entre la partición de datos de la muestra utilizada para el entrenamiento del modelo y la partición de datos utilizada para su validación. Se basa en dividir los datos

aleatoriamente en k grupos, dejando un conjunto i a parte y construyendo el modelo con el resto de grupos (k-i). Finalmente se estima el error con el set i. Este proceso lo repetimos el número de veces que especifiquemos cambiando la semilla de aleatorización del proceso. En este trabajo establecemos por defecto el uso de un número de 4 grupos y la ejecución de 20 semillas diferentes cada vez.

3 Selección inicial de la muestra de datos

La empresa facilita de forma pública en su portal web ficheros en formato *csv* en cuatrimestres desde sus inicios. En estos archivos se incluye información asociada al solicitante del préstamo, desde información externa de burós de crédito a información relacionada con su historial de pagos o al propio préstamo solicitado.

Hemos considerado apropiado comenzar a trabajar con un set de datos correspondiente a los préstamos concedidos en Q2 del 2018 (abril, mayo y junio), con un total de 130.773 registros y 144 variables. Se ha seleccionado este periodo ya que es el más adecuado para realizar este tipo de análisis, dado que Q1 presentaría resultados algo más alejados de la realidad dada la tendencia a observar mayor dificultad financiera de los individuos durante los primeros meses del año. Además, 2018 es un buen año para llevar a cabo el análisis dado que la economía ha medrado con respecto a su situación inestable que trajeron los duros años de la crisis financiera, evitando así obtener tasas de impago demasiado elevadas que se alejen de la realidad.

3.1 Clasificación de las variables

El set inicial de datos contiene un total de 144 variables las cuales pueden clasificarse de la siguiente forma:

- Variables relacionadas con la solicitud del préstamo (3). Pueden estar relacionadas con la identificación (1) y con el importe solicitado (2).
- Variables relacionadas con el solicitante (89). En esta categoría encontramos lo siguiente:
    - Variable identificadora (1).
    - Variables de caracterización social (6).
    - Variables de caracterización económica (5).
    - Variables que desvelan indicios de impago del historial crediticio (14).
    - Variables que desvelan información de procedencia externa sobre todas las cuentas de las que dispone o ha dispuesto el individuo y su caracterización (34).
    - Variable que desvela las consultas al registro público y sus resultados (7).
    - Variable que desvela información del estado de balance de las diferentes cuentas existentes (20).
    - Variable que desvela información de ratings externos (2).
- Variables relacionadas con el préstamo actual con LendingClub (52):
    - Variables que desvelan información sobre el balance (4).
    - Variables que definen y caracterizan el préstamo (9).
    - Variables con información sobre los pagos que debe afrontar y su estado (10).
    - Variables que reflejan información sobre la calificación del préstamo hecha por LendingClub para sus inversores (2).
    - Variable que refleja el estado del préstamo (1).

- Variables que aplican solamente a aquellos casos que han requerido plan de ayuda de reestructuración de deuda (16).
- Variables relacionadas con planes de recobro del préstamo (3).
- Variables que aplican solo cuando se ha llevado a cabo una colaboración con una empresa externa para conseguir liquidar la deuda (7).

La descripción detallada de las variables del set inicial se encuentra en el Anexo I.

3.2 Definición de la variable objetivo

La finalidad del trabajo es realizar un modelo de detección de impago en el momento de solicitud de préstamo para, a partir de las características de cada individuo, poder detectar aquellos con mayor probabilidad de impago y evitar concederles un préstamo. Se trata de analizar y extraer la tendencia de comportamiento según sean las características observadas en cada una de las variables input para aquellos que ya son clientes y, por tanto, ya hemos podido evaluar su comportamiento. El siguiente paso es conformar un modelo con toda esta información de los clientes actuales ya disponible en el momento de solicitud como input para utilizarlo como mecanismo de prevención de impago y pérdida futura. Por ello, es necesario identificar en nuestro set de datos una variable que revele el comportamiento relacionado con los pagos de los individuos a los que se les ha concedido ya algún préstamo. De esta forma seleccionamos como variable más adecuada la que refleja el estado del préstamo ***loan_status***, en la que se identifican estas categorías a partir de las cuales construimos la variable objetivo como se muestra a continuación:

Figura 3: Definición Variable Objetivo Default

loan_status	Default
Current	0= no
Fully Paid	0= no
In Grace Period	0= no
Late (16-30 days)	1= yes
Late (31-120 days)	1= yes
Default	1= yes
Charged Off	1= yes

Cabe realizar una breve explicación de qué significa cada una de las categorías de la variable para comprender la razón por la cual consideramos que esta es la reagrupación adecuada de las mismas para componer nuestra variable objetivo dicotómica:

- *Current*: estado de los pagos al corriente, se paga dentro de los plazos estipulados.
- *Fully paid*: préstamo que ya se ha pagado al completo.
- *In grace period*: préstamo fuera de plazo de pago pero dentro de los 15 días de gracia.
- *Late (16-30 days*): préstamo fuera de plazo de pago por más de 16 días y menos de 30.
- *Late (31-120 days*): préstamo fuera de plazo por más de 31 días y menos de 120.
- *Default*: préstamo impagado por un periodo extenso de días superior a 120, pero que todavía se tiene esperanza de que se pueda ejecutar recobro.
- *Charge*-off: préstamo impagado por largo periodo de tiempo sin esperanza de recobro.

Dicho esto, claramente las categorías *Current* y *Fully Paid* reflejan préstamos sin ningún problema de impago, por lo que son consideradas no *default* (0). Por otra parte, la variable *in grace period* no está tan claro que se vaya a recuperar el pago, pero no se considerará impago por ahora. Las categorías restantes son aquellas que consideramos categorizar

como préstamos impagados: *Late (16-30 days), Late (31-120 days), Default* y *Charge-off*. A pesar de ello, es importante tener en cuenta que podría cambiar el estado y un préstamo *Late* o incluso un *Default* podrían ser recuperados. Como no contamos con esta información, por ahora simplemente nos basamos en llevar a cabo un análisis simple con la información disponible en el momento actual y sin considerar la evolución de los estados a medida que evoluciona la ventana temporal. Así entonces conseguimos nuestra variable objetivo ***DEFAULT***, la cual tomará valor 0 si no es casuística impago y valor 1 si lo es.

3.3 Modificación y preparación del set de trabajo

En este apartado trataremos de plasmar todo el trabajo previo a la modelización que se ha tenido que llevar a cabo para conseguir un set de datos con el que realmente se pueda trabajar alineadamente con los objetivos fijados.

3.3.1 Descarte Previo de Variables

Dada la extensión de variables contenidas en nuestro set de datos hemos tenido que realizar una exploración inicial exhaustiva para familiarizarnos con el contenido y con su significado. Tras el análisis inicial extraemos una clara necesidad de simplificación del set de datos para que sea posible trabajar con ellos.

3.3.1.1 Variables Restringidas

Lo primero que hemos identificado ha sido la existencia de información solamente disponible para descargas desde cuentas de inversores registrados en la plataforma con finalidad de invertir. Se trata principalmente de información identificativa e interna, la cual ha sido solicitada a la empresa, pero finalmente no hemos podido conseguirla. Por ello obviaremos estas variables y estableceremos nuestra propia numeración identificativa.

3.3.1.2 Variables unarias e inutilizables

Además, existen algunas otras variables vacías o unarias y otras variables de texto descriptivo pero sin patrones suficientes para realizar *text mining* (muchos registros vacíos y con carencia de estructura). Todas estas variables o no pueden usarse o complican de forma innecesaria el estudio, por lo que hemos decidido prescindir de ellas.

3.3.1.3 Fusión de variables con información paralela

Otra tarea de simplificación importante que hemos llevado a cabo ha sido la fusión de variables con información desagregada. A priori hemos visto que tenemos dos tipos de solicitud de préstamo dependiendo de si se realizan individual (un solo prestatario) o conjuntamente (dos co-prestatarios). Esto da lugar a que existan variables que solamente tomen valor si la solicitud es conjunta ya que hacen referencia al co-prestatario y están en blanco para las solicitudes individuales. Lo que hemos hecho finalmente ha sido analizar en cada caso qué tipo de variable tenemos para extraer la mejor forma de agregar en una sola columna ambas. En algunos casos se ha realizado una suma (variables que reflejan el número total de cuentas, renta, balance etc.), en otros casos se ha hecho la media (ratios) y en otros casos no se ha podido agregar ya que hacerlo no suponía recoger información realista y hemos mantenido solamente la columna con información individual.

3.3.1.4 Variables con aporte mínimo de información

Las variables que hacen referencia a planes de ayuda frente a dificultad de pago de deuda, de reestructuración de deuda por imposibilidad de pago del préstamo y de información de colaboración con empresas externas de liquidación de deuda, solamente están disponibles para aquellos registros que ya cuentan con un préstamo fallido. Entonces, concluimos que no podemos incluir todas estas variables en el estudio debido a que contienen muchos datos faltantes y no procede su utilización por motivos de temporalidad. Si realizáramos un análisis más en profundidad y con otra ventana temporal, centrándonos en los registros con impago sí podrían ser de ayuda, pero no se trata del objetivo de este trabajo y por eso decidimos prescindir de ellas.

3.3.1.5 Variables futuras desconocidas en el momento de solicitud

Si recapitulamos y tratamos de volver a situar en el centro el objetivo del trabajo, vemos que queremos crear una herramienta que le permita a LendingClub identificar individuos con mayor probabilidad de impago para evitar concederles un préstamo. Por ello, es importante que la construcción del modelo solo cuente con variables input que se conocen en el momento que el individuo realiza la solicitud, es decir, que no sean generadas a partir de la evolución de su comportamiento una vez ya hemos concedido el préstamo. Comprobamos además que estas variables generadas a posteriori no nos servirán para modelizar llevando a cabo una prueba de aporte con la que vemos que, efectivamente, presentan una influencia anómalamente elevada en comparación al resto debido a que es como si estuviéramos utilizando la propia palabra definida para llevar a cabo la definición. Por ello, para este estudio nos deshacemos en este momento de todas las variables generadas en un momento temporal posterior al de la solicitud del préstamo.

3.3.1.6 Tratamiento de variables con datos faltantes

En el set de datos hemos detectado algunas variables con datos vacíos pero no pueden ser considerados *missing* debido a que sería como una categoría en la que se identificarían aquellos registros para los que no procede dar valor en esa característica. Un ejemplo es cuando se habla de una variable que refleja el número de tarjetas de crédito activas o el balance total en tarjeta de crédito; si un individuo no cuenta con ninguna tarjeta contará con dato nulo en esta variable, pero se trata de un dato nulo representativo que aporta información. El problema detectado es que en la mayoría de casos se trata de variables de intervalo, por lo que la única forma de mantenerlas con esos datos faltantes y que a la vez sean compatibles con todos los modelos es convirtiéndolas a categóricas para poder mantener el máximo de información posible.

En el caso de los ratios para los que el denominador es cero, como no puede ser calculado aparece vacío el campo, lo cual no es lo mismo que tomar valor cero. Por ejemplo, el ratio de límite de endeudamiento de tarjetas de crédito, si no se cuenta con cuenta con ninguna tarjeta, no tenemos valor y aparece como *missing,* lo cual no es lo mismo que realmente la división valga 0. Es muy importante analizar y estudiar qué información recogen todas y cada una de las variables para evitar incurrir en suposiciones erróneas o tratamiento inadecuado, por lo que decidimos que en este caso no es correcto asumir que los nulos equivalen a cero y hemos recategorizado estas variables para ver si pueden aportar información en este formato antes que descartarlas.

3.3.1.7 Eliminación de registros con muchos datos faltantes

Durante el análisis exploratorio de las variables se han detectado siete registros con *missing* en más del 90% de las variables del set. Parece tratarse de préstamos que han sido solicitados y aceptados pero finalmente no han llegado a materializarse. Por ello, como estos registros no aportan nada es mejor eliminarlos del estudio.

3.3.2 Creación de nuevas variables

Por otra parte, también se ha analizado el potencial de creación de nuevas variables, llevándonos a generar, a partir de ciertas variables existentes, otras que podrían aportar al modelo:

Figura 4: Creación de Nuevas Variables

Nueva variable	Descripción
mths_earliest_cr_line	Meses transcurridos desde la apertura de la primera línea de crédito hasta la fecha de solicitud de crédito
nom_revol_bal	Balance "revolving" disponible / renta disponible
nom_tot_cur_bal	Balance total de todas las cuentas / renta disponible
nom_total_bal_il	Balance total cuentas "installment" / renta disponible
nom_max_bal_bc	Balance máximo que se debe en las cuentas "revolving"/ renta disponible
nom_avg_cur_bal	Balance medio actual de todas las cuentas / renta disponible
pct_bc_sats	Nº cuentas bancarias satisfactorias / total cuentas
pct_all_sats	Nº cuentas satisfactorias / total cuentas
nom_tot_bal_ex_mort	Balance total (excluyendo hipoteca) / renta disponible
delinq_account	Variable dicotómica que refleja el historial de delincuencias del principal

Todas aquellas variables con el prefijo ***nom*** se crean a partir de la división la variable original entre la renta disponible para así conseguir su normalización. Lo que haremos para estas variables es ver si aportan más en este formato o el original cuando realicemos las pruebas de aporte. La variable ***mths_earliest_cr_line*** ha sido creada a partir de una variable que incluía muchas fechas de apertura de líneas de crédito, lo cual la hacía categórica con demasiadas clases. Hemos calculado el número de meses transcurridos desde la fecha de apertura hasta la fecha de hoy para cada registro, dando lugar así a una variable de intervalo. Para las variables ***pct_bc_sats*** y ***pct_all_sats*** simplemente hemos normalizado dividiendo el número de cuentas satisfactorias entre el total de cuentas para obtener una variable más representativa. Finalmente creamos la variable ***delinq_account***, la cual es dicotómica y toma valor 1 si el solicitante principal tiene cuentas con delincuencia y 0 si no. Esta variable la creamos ya que no disponemos de una variable que haga referencia a las cuentas delincuentes como tal, solamente tenemos variables de meses transcurridos desde la última delincuencia o similares creamos esta variable por si pudiera ser finalmente útil.

3.3.3 Análisis factorial para simplificar la matriz definitiva

La última etapa del proceso de simplificación ha sido realizar un análisis factorial para las variables de intervalo del set utilizando componentes con rotación *Varimax*. Esta técnica estadística se utiliza para llevar a cabo la reducción de los datos basándose en explicar las correlaciones entre las variables utilizando otras variables ficticias generadas, los factores. Se crean entonces tantos factores como grupos de variables que representan información diferente dentro del set de datos. Nosotros simplemente usamos estos factores para detectar los grupos de información que tenemos en el set de datos y para cada grupo

mantener aquellas variables más relevantes y deshacernos de las demás, evitando solapamiento y redundancia. De esta forma además evitamos introducir demasiadas variables para prevenir la sobre parametrización de los modelos. El resultado de este proceso es el siguiente set con 39 variables:

Figura 5: Tabla Variables Definitivas Estudio

Variable	Descripción
id	Identificador
loan_amnt	Cantidad de crédito solicitada
term	Número de pagos (en meses) elegidos para devolver el crédito: toma valores 36 o 60 meses
installment	Cantidad mensual a pagar por el crédito
emp_length	Tiempo que lleva trabajando el individuo: <1, 1, 2 … 9, 10+, n/a
home_ownership	Caracterización del status del individuo respecto a su vivienda actual: RENT, OWN, MORTGAGE, ANY
annual_inc_both	Renta anual reportada por el / los solicitante(s)
verification_status_joint	Estado de verificación de la renta: puede ser no verificado, verificado, verificada la fuente de ingreso
DEFAULT	VARIABLE OBJETIVO. Construida a partir de una variable que indica el estado de los préstamos en la actualidad.
purpose	Uso que se le dará al crédito: coche, deuda de tarjetas, consolidación de deuda, reforma, casa, compra mayor, medicina, mudanza, energía renovable, otros
addr_state	Estado del solicitante (50 estados de Estados Unidos de América)
dti_total	Ratio: pago mensual de deuda dadas las obligaciones (excluyendo el monto de la hipoteca y el crédito solicitado a LC) / renta mensual reportada
delinq_2yrs	Número de vencimientos a + de 30 días (delincuencias) que aparecen en el registro dentro de los últimos 2 años
mths_earliest_cr_line	Meses transcurridos desde la apertura de la primera línea de crédito hasta la fecha de solicitud
both_inq_last_6mths	Número de peticiones de revisión de historial de crédito realizadas en los dos últimos meses (suma de los dos individuos en caso que sea aplicación conjunta)
nom_revol_bal	Balance "revolving" disponible / renta disponible
total_acc	Total líneas de crédito contenidas en el archivo de crédito
application_type	Tipo de aplicación: individual o conjunta (variable referencia*)
nom_tot_cur_bal	Balance total de todas las cuentas / renta disponible
open_il_24m	Cuentas de tipo "installment" abiertas durante los últimos 24 meses
nom_total_bal_il	Balance total cuentas "installment" / renta disponible
il_util	ratio límite de crédito / balance nº de cuentas de tipo "installment"
open_rv_12m	Cuentas de tipo "revolving" abiertas en los pasados 12 meses
all_util	Ratio: balance total del límite de crédito / balance de todas las cuentas
total_rev_hi_lim	Balance total de límite de crédito en cuentas "revolving"
inq_fi	Número de consultas financieras personales efectuadas para el individuo
acc_open_past_24mths	Número de cuentas abiertas durante los últimos 24 meses
both_mort_acc	Suma del número de hipotecas total
mths_since_recent_inq	Meses transcurridos desde la consulta pública realizada más recientemente
num_rev_tl_bal_gt_0	Número de cuentas tipo "revolving" con balance >0
pct_all_sats	Número cuentas satisfactorias / total cuentas
pct_tl_nvr_dlq	Porcentaje de cuentas en las que nunca han sido delincuentes
pub_rec_bankruptcies	Número de veces que ha estado en bancarrota constatadas en el registro público
tax_liens	Número de embargos fiscales constatados en el registro público
nom_tot_bal_ex_mort	Balance total (excluyendo hipoteca) / renta disponible
disbursement_method	Método de pago por el cual se recibe el crédito: cash o direct pay
tot_hi_cred_lim	Balance total de límite de crédito en todas las cuentas
total_bc_limit	Balance total de límite de crédito en tarjetas de crédito
both_chargeoff_12m	Suma del número de cuentas declaradas incobrables de los últimos 12 meses

Cabe mencionar que, para seleccionar qué variable era más conveniente mantener no solo nos hemos fijado en la matriz del análisis factorial y en su porcentaje de aporte al factor, sino que también hemos teniendo en cuenta diversos análisis realizados en SAS Miner en cuanto a la posición y el aporte de cada variable utilizando el estadístico Valor. Además, un detalle que tenemos en cuenta es especificar qué variables son de nueva creación: aquellas marcadas en azul claro surgen tras realizar la unión de dos variables existentes, mientras que las marcadas en azul oscuro surgen tras realizar alguna modificación a variables ya existentes (normalización, creación de nuevas variables con aporte de información diferente en base a la original etc). Hemos identificado que aportaban más que las no transformadas con respecto a las de su mismo factor, con información similar.

3.4 Aclaración relacionada con los diferentes tipos de cuentas crediticias

Como bien se ha comentado, el procedimiento de exploración de los datos ha requerido mucho tiempo y realizar un análisis profundo debido a la gran cantidad de variables con significado semejante, pero con pequeños matices diferentes que hemos tenido que investigar y considerar para determinar la utilidad de las variables. Además, debido al uso de nomenclatura técnica y de las explicaciones algo confusas facilitadas por parte de la empresa sobre las variables, hemos tenido que analizar de forma individual todas y cada una de ellas y extraer el significado completo. Este procedimiento ha requerido indagar en conceptualización bancaria y crediticia para poder valorar objetivamente el contenido y evaluar la utilidad de la información. Dicho esto, consideramos oportuno realizar algunas aclaraciones con respecto a los diferentes tipos de cuenta bancaria a los que alude la información contenida en el set de datos:

- ***Revolving accounts***: se trata de cuentas que contienen un balance que fluctúa mes a mes dependiendo de la cantidad que se utiliza. Esta categoría está compuesta principalmente por tarjetas de crédito, aunque también puede incluir líneas de equidad.
- ***Installment accounts***: cuentas que cuentan con cantidad y fecha de fin prefijada. Se pacta a priori la planificación de amortización y se sabe desde el primer momento como se debería ir reduciendo el principal a medida que se van efectuando los pagos correspondientes a lo largo del tiempo. Ejemplos de este tipo de cuentas son las hipotecas y créditos.

Es importante distinguir entre estos tipos de cuentas ya que, dependiendo de la composición de la deuda del individuo en unos u otros, tendrán mayor o menor riesgo de impago. En los datos se incluye información sobre los dos tipos de cuenta principal (*revolving* e *installment*) y dentro de estas de las tarjetas de crédito (*bank card*) e hipotecas (*mortgage*). Además, como no hay una palabra en español para hacer referencia a *revolving* o *installment*, se sigue utilizando estos dos términos anglosajones en el trabajo.

3.5 Sobre muestreo

Se ha tomado la decisión de realizar una submuestra de los datos obtenida mediante el sobre muestreo para realizar el trabajo. Para explicar en qué consiste este proceso, primero de todo, debemos recordar que el objetivo es detectar los impagos (casuística en la variable objetivo *defaul*t = 1) y mencionar entonces que inicialmente contamos con un 4.2% de esta casuística (lo cual supone un total de 5.451 individuos) frente al 95.8% restante de no *default=0*. Lo que se ha hecho entonces es crear un conjunto de datos con

todas esas casuísticas default = 1 y se ha realizado una muestra aleatoria de las casuísticas contrarias default = 0 de 5.452 observaciones también. De esta forma, creamos una columna de frecuencia para poder ponderar posteriormente los errores para saber que por cada uno de estos casos default, en realidad, existen 22,99[1] casos de default= 0 (aunque en la muestra se observe una casuística 50/50). Este método de selección muestral nos permitirá superar las limitaciones computacionales que suponía trabajar con el set de datos inicial. En este trabajo, por cuestiones de tiempo y espacio modelizaremos únicamente para un set de datos, pero luego podría ser interesante probar la estabilidad de los datos y de los modelos realizando diferentes sets de datos sobre muestreados variando aleatoriamente los casos default=0 y ver si los resultados de los modelos cambian según el set de datos utilizado.

Finalmente, tras haber realizado todas estas modificaciones de preparación previa y simplificación del set de datos contamos con un total de 10.902 registros y 39 variables como punto de partida para la exploración y la depuración.

4 Depuración y análisis descriptivo exploratorio del set de datos

En este apartado llevaremos a cabo un análisis descriptivo y exploratorio de los datos para conocer la información que contienen, su formato, medidas, contenido de *missings* y datos atípicos que puedan generar problemas en nuestra modelización. Además, comprobaremos si existiera alguna transformación que pudiéramos aplicar a alguna variable para mejorar la relación con la variable objetivo. A priori, el set contiene 39 variables de las cuales tenemos 1 variable identificadora, la variable objetivo binaria y 10 categóricas, de las cuales 3 son binarias y 28 variables de intervalo.

4.1 Análisis univariante

A continuación, vamos a explorar nuestras variables para tener una foto inicial de las respuestas, valores y categorías que presentan los individuos en cada una de las variables separadas según sean de intervalo o categóricas.

4.1.1 Variables categóricas

Figura 6:Tabla Exploración Variables Categóricas

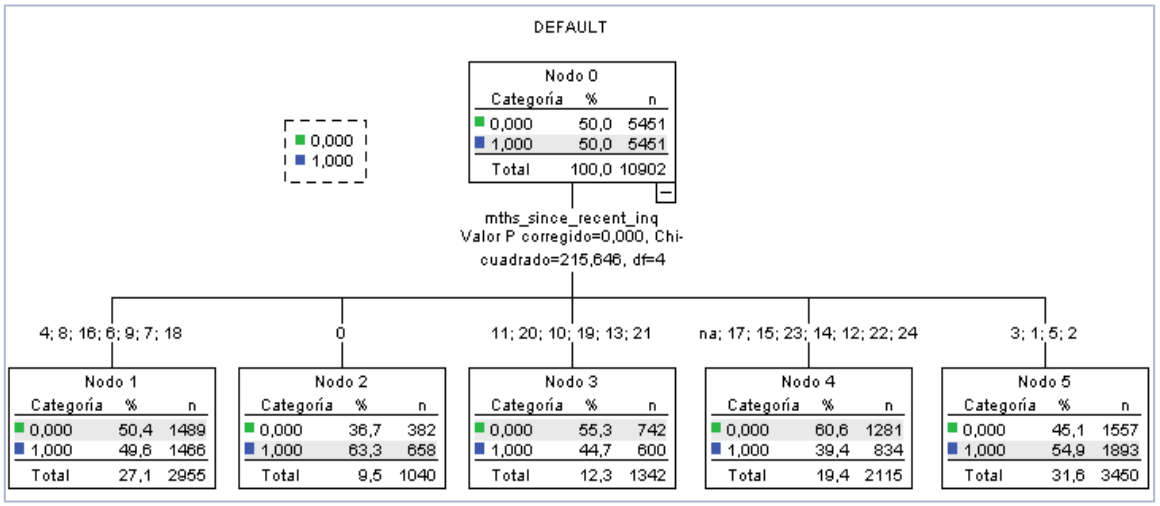
Variable	Tipo	Niveles	Ausentes
DEFAULT	N	2	0
addr_state	N	4	0
application_type	C	2	0
disbursement_method	C	2	0
emp_length	C	12	0
home_ownership	C	4	0
mths_since_recent_inq	N	5	0
purpose	C	12	0
term	N	2	0
verification_status_joint	C	3	0

Como podemos observar para las 10 variables categóricas con las que se cuenta, no tenemos datos ausentes. Todas presentan cantidades apropiadas de niveles, excepto

[1] Este factor ponderación se calcula dividiendo el total de casos de default=0, es decir, las 125.314 observaciones con ese comportamiento entre las de default = 1, que son 5.451.

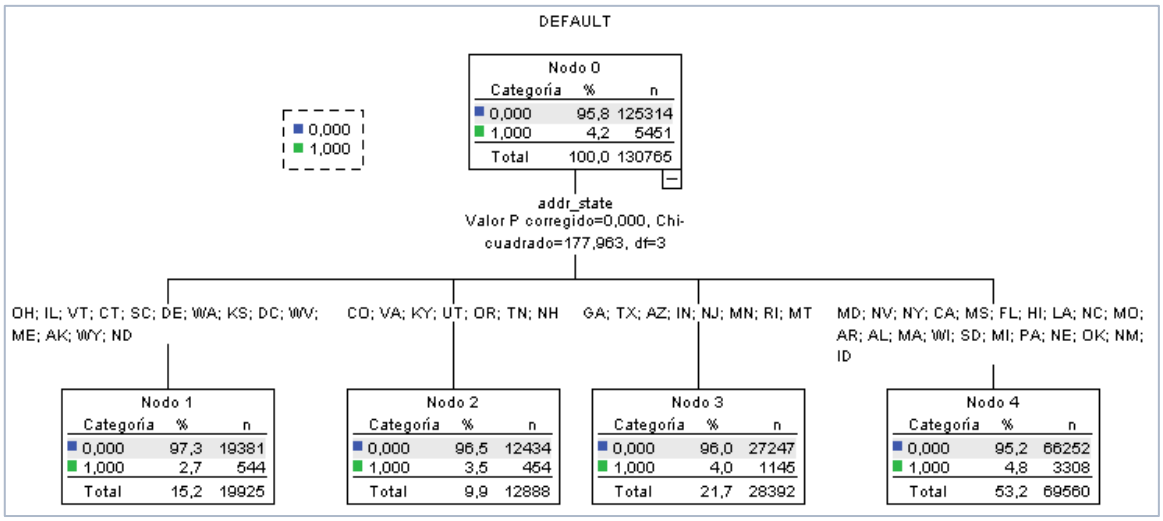
mths_since_rcnt_inq y ***addr_state***. Para estas, como tenemos demasiadas categorías y esto supone que algunas contengan muy pocas observaciones, hemos utilizado un árbol de decisión para reagruparlas de la siguiente forma:

Figura 7: Árbol para reagrupar la variable mths_since_rcnt_inq



El árbol desvela que la mejor forma de agrupar las categorías de ***mths_since_rcnt_inq*** manteniendo coherencia y relación de las categorías recogidas con respecto la variable objetivo es la siguiente: si hace 0 meses que se ha revisado el historial de crédito (valor 0), el nivel de riesgo de impago es más elevado, por lo que se mantiene en una categoría sola, la categoría 0. A continuación, si hace entre 1 y 5 meses que te han revisado, disminuye algo el riesgo y se crea la categoría 1 (aunque 4 meses el árbol los mete en otra categoría, los meteremos aquí para evitar dar un salto y seguir el orden). La categoría es la 2, con individuos a los que no se ha revisado desde hace entre 6 y 9 meses (incluyendo también entre 16 y 18 meses). En la categoría 3 se incluyen los que no se han revisado desde hace entre 10-20 y finalmente el grupo 4 con más de 20 meses y los *na* (no aplica), representando los individuos que no han sido revisados nunca.

Figura 8: Árbol para reagrupar la variable addr_state



Para los 50 Estados de EEUUA hemos optado por la agrupación que nos da el árbol anterior. De esta forma conseguimos cuatro grupos con diferencias en el porcentaje de impago.

4.1.2 *Variables de intervalo*

Figura 9: Tabla Exploración Variables de intervalo

Variable	Ausentes	Mediana	Mínimo	Máximo	Media	Desv. Est	Asimetría	Curtosis
acc_open_past_24mths	0	4.00	0	54	4.53	3.31	1.45	4.2
all_util	35	55.00	0	156	54.07	21.32	-0.1	-0.18
annual_inc_both	0	73000.00	2300	9930475	87704.26	91539.5	43.38	3686.91
both_mort_acc	0	1.00	0	47	1.51	2.06	2.25	9.49
delinq_2yrs	0	0.00	0	23	0.24	0.75	6.49	76.24
dti_total	0	17.03	0	39.99	17.61	8.71	0.27	-0.54
il_util	0	67.00	0	619.207037	60.19	32.14	-0.44	0.82
inq_fi	0	1.00	0	31	1.11	1.55	2.7	14.67
installment	0	381.74	30.12	1628.08	463.11	286.2	0.94	0.32
loan_amnt	0	13500.00	1000	40000	15908.41	10129.77	0.78	-0.28
mths_earliest_cr_line	0	172.00	37	761	191.26	94.23	1.1	1.76
nom_revol_bal	0	16.64	0	647.442791	20.77	18.77	4.17	57.53
nom_tot_bal_ex_mort	0	49.60	0	7576.9	63.29	64.95	16.79	1449.05
nom_tot_cur_bal	0	99.66	0	7576.9	157.21	149.16	2.34	50.14
nom_total_bal_il	0	29.21	0	7520.4	44.15	62.23	18.3	1667.69
num_rev_tl_bal_gt_0	0	5.00	0	65	5.27	3.38	1.57	5.22
open_il_24m	0	1.00	0	18	1.52	1.57	1.67	4.54
open_rv_12m	0	1.00	0	26	1.28	1.54	2.04	7.6
pct_all_sats	0	53.85	0	100	55.2	18.94	0.27	-0.38
pct_tl_nvr_dlq	0	100.00	9.1	100	94.43	9.27	-2.43	7.58
tax_liens	0	0.00	0	15	0.01	0.15	29.14	1417.63
tot_hi_cred_lim	0	112762.00	0	5850873	180019.42	182545.28	2.68	23.31
total_acc	0	20.00	2	140	22.47	12.16	1.15	2.3
total_bc_limit	0	18100.00	0	1569000	25424.31	25326.94	4.33	119.78
total_rev_hi_lim	0	27500.00	0	2087500	36962.72	37848.11	8.95	293.57
both_chargeoff_12m	0	0.00	0	20	0.01	0.18	45.67	3815.66
pub_rec_bankruptcies	0	0.00	0	7	0.12	0.34	2.61	7.70
both_inq_last_6mths	0	0.00	0	9	0.54	0.87	2.18	7.03

Lo primero que analizamos es el número de datos *missing* de las variables. A priori, ya no hemos metido en el estudio inicial aquellas variables que pudieran tener un número elevado que pudiera generar problemas en el estudio, por lo que identificamos solamente una variable con datos faltantes. Además, cuando hemos tenido que elegir en el análisis factorial entre dos variables que prácticamente aportaban lo mismo también hemos tenido en cuenta los datos faltantes ya que, al tratarlos, tanto si se eliminan las observaciones como si se imputan, estamos o bien perdiendo información o bien sesgando la realidad, por lo que es preferible usar variables sin necesidad de tratamiento de atípicos. Además, hay variables que por definición es correcto que contengan datos *missing,* ya que ese hueco contiene información implícita como ya hemos mencionado anteriormente. Por ello, hemos tenido que realizar un trabajo muy manual analizando caso por caso y según la información recogida si procedía recategorizarla o no para poder utilizar la variable como input. Un ejemplo de ello ha sido ***mths_since_rcnt_inq***, para la hemos rellenado los datos *missing* con *na,* reflejando aquellas solicitudes para las que nunca ha sido necesario realizar consultas al registro público de deuda. Para las variables que finalmente usamos en el estudio ninguna presenta un número de datos *missing* importante, por lo que simplemente imputaremos la variable ***all_util*** como explicaremos posteriormente.

A continuación, nos aseguramos de que las variables no contienen números incorrectos (negativos o absurdos en relación a su significado) y confirmamos así que todo es correcto. Nos fijamos también en la desviación estándar para tener una idea inicial de cómo se alejan ciertas observaciones de la media para cada variable, indicador que puede permitir identificar datos atípicos. También nos fijemos en los niveles de asimetría y curtosis; la asimetría suele venir principalmente generada por datos atípicos que alteran la distribución simétrica de las variables, por lo que podemos definir como asimétricas aquellas con valores alejados de -1, 0 y 1. Además, la curtosis también se ve alterada cuando la variable contiene alguna agrupación anómala de las observaciones. En la figura 9 de la página anterior marcamos en amarillo todas las variables para las que conviene revisar el contenido de datos atípicos y ver en cada caso si procede tratarlos.

4.1.3 Tratamiento de datos atípicos

En este apartado vamos a centrarnos en la detección de datos atípicos que pudieran desviar la representatividad de nuestros modelos. Si bien es cierto que los datos atípicos pueden ser problemáticos, también puede ser un inconveniente adoptar una posición demasiado restrictiva y eliminar información de más. Por ello, utilizaremos el consenso de dos de los siguientes métodos para marcar los límites a partir de los cuales procede considerar un registro como dato atípico. El primero de los dos métodos depende de la caracterización estadística de nuestras variables; por ello, primero nos fijamos en cómo es la distribución de las variables: si esta es simétrica utilizamos como primer método el de la desviación estándar, mientras que si son asimétricas utilizamos el método de la desviación absoluta de la media. En caso de que sean asimétricas y además contengan mediana igual a cero, utilizaremos el método de percentiles extremos. El segundo método que usamos para determinar la atipicidad es el de rango intercuartílico, el cual se utiliza para todas las variables, independientemente de sus características. Se lleva a cabo este proceso de determinación de límites superior e inferior para cada variable a partir de los cuales se debería considerar como atípico el dato; para ello se seleccionan de los límites obtenidos con cada método, el umbral superior e inferior menos restrictivo para evitar suprimir información de más. En el anexo III se adjunta la tabla detallada con cuáles son los límites obtenidos para cada variable por cada uno de los métodos, el límite generado por consenso y la cantidad de datos atípicos identificados en cada caso.

Finalmente, como no se detecta una cantidad de datos atípicos lo suficientemente grande como para dejar de contar con alguna variable, consideramos oportuno utilizar el método de imputación para tratarlos. Este método consiste en convertir a *missing* estos datos e imputarlos utilizando un árbol para que tomen valores reales y parecidos a los que toman otras observaciones con características semejantes. De esta forma también se imputa la variable ***all_util*** que hemos especificado en el apartado de datos faltantes.

4.2 Análisis Multivariante

4.2.1 Transformación de Variables

Dada la complejidad que a priori transmite el contenido de las variables del estudio, puede ser interesante tratar de plasmar mejor la relación en cuanto a la variable objetivo o de generar información adicional mediante el uso de transformaciones. Es importante ver si realmente existe alguna relación no lineal que no se pueda percibir a simple vista y pueda aportar información adicional. Para ello utilizamos el nodo transformación de variables en SAS Miner y establecemos el criterio "mejor", la cual transforma las variables de forma que se mejore la relación con la variable objetivo para las variables de intervalo. De esta forma, añadimos a la matriz de datos las siguientes transformaciones y dejaremos que sean los propios métodos de selección los que determinen si generan aporte y si procede que las mantengamos para la modelización.

Figura 10: Tabla de Variables Creadas con el nodo Transformación de Variables

Nombre Variable	Fórmula	Niveles
EXP_pct_all_sats	exp(pct_all_sats)	Intervalo
INV_both_mort_acc	1 / (both_mort_acc + 1)	Intervalo
INV_num_rev_tl_bal_gt_0	1 / (num_rev_tl_bal_gt_0 + 1)	Intervalo
INV_pub_rec_bankruptcies	1 / (pub_rec_bankruptcies + 1)	Intervalo
INV_total_acc	1 / (total_acc + 1)	Intervalo
SQRT_acc_open_past_24mths	Sqrt(acc_open_past_24mths + 1)	Intervalo
SQRT_loan_amnt	Sqrt(loan_amnt + 1)	Intervalo
SQRT_open_rv_12m	Sqrt(open_rv_12m + 1)	Intervalo
SQR_pct_tl_nvr_dlq	(pct_tl_nvr_dlq + 1)**2	Intervalo

4.2.2 Pruebas de aporte: test de Welch y Chi-cuadrado

En este apartado buscamos llevar a cabo una primera identificación de aquellas variables que pueden ser útiles para el objetivo de detección de solicitudes de crédito malas (con mayor probabilidad de impago). Para ello, dado que en los gráficos es algo complejo identificar la capacidad discriminatoria de las variables, llevamos a cabo el test de medias de Welch para las variables de intervalo y el test Chi-cuadrado para las variables categóricas. De esta forma consideraremos la opción de prescindir de las variables no significativas para la modelización. En la figura 11 adjuntamos una tabla resumen de los resultados para cada variable.

Tras analizar el p-valor a un nivel de significación del 5%, todas las variables marcadas en verde aportan en el ejercicio de discriminación de solicitudes buenas (no impago) y malas (impago), mientras que las seis marcadas en rojo las podrían ser descartadas por ser no significativas.

Figura 11: Resumen de las pruebas de aporte de las variables

Variable	Tipo	Estadístico	Resultado	P-valor	Aporte[2]
addr_state	categórica	Chi-cuadrado	108.87	1.92438E-23	si
application_type	categórica	Chi-cuadrado	13.295	0.0002661	si
disbursement_method	categórica	Chi-cuadrado	116.27	4.15154E-27	si
emp_length	categórica	Chi-cuadrado	89.01	2.605E-14	si
home_ownership	categórica	Chi-cuadrado	106.98	4.88835E-23	si
mths_since_recent_inq	categórica	Chi-cuadrado	204.05	5.05556E-43	si
purpose	categórica	Chi-cuadrado	149.1	2.28E-26	si
term	categórica	Chi-cuadrado	81.639	1.63E-19	si
verification_status_joint	categórica	Chi-cuadrado	197.464	1.32E-43	si
loan_amnt	intervalo	Welch	54.09742	2.05E-13	si
installment	intervalo	Welch	98.73561	3.62E-23	si
dti_total	intervalo	Welch	1.43233	0.2314108	no
both_inq_last_6mths	intervalo	Welch	206.706	1.99441E-46	si
pct_all_sats	intervalo	Welch	0.009396757	0.9227783	no
pct_tl_nvr_dlq	intervalo	Welch	6.151186	0.01314742	si*
pub_rec_bankruptcies	intervalo	Welch	5.391452	0.02025399	si*
tax_liens	intervalo	Welch	1.734692	0.1878431	no
both_chargeoff_12m	intervalo	Welch	9.233532	0.002385439	si
acc_open_past_24mths	intervalo	Welch	93.64076	4.64646E-22	si
annual_inc_both	intervalo	Welch	22.13352	2.57465E-06	si
both_mort_acc	intervalo	Welch	50.51252	1.25839E-12	si
delinq_2yrs	intervalo	Welch	7.269849	0.007023026	si
il_util	intervalo	Welch	13.32175	0.000263573	si
inq_fi	intervalo	Welch	77.88116	1.26584E-18	si
mths_earliest_cr_line	intervalo	Welch	41.55095	1.19691E-10	si
nom_revol_bal	intervalo	Welch	4.009105	0.04527993	si*
nom_tot_bal_ex_mort	intervalo	Welch	1.090618	0.2963574	no
nom_tot_cur_bal	intervalo	Welch	42.985	5.76553E-11	si
nom_total_bal_il	intervalo	Welch	1.157837	0.2819383	no
num_rev_tl_bal_gt_0	intervalo	Welch	9.256449	0.00235226	si
open_il_24m	intervalo	Welch	49.92776	1.69353E-12	si
open_rv_12m	intervalo	Welch	48.80952	2.98664E-12	si
tot_hi_cred_lim	intervalo	Welch	97.59328	6.42274E-23	si
total_acc	intervalo	Welch	13.60921	0.000226174	si
total_bc_limit	intervalo	Welch	94.16178	3.57976E-22	si
total_rev_hi_lim	intervalo	Welch	87.49448	1.0113E-20	si
all_util	intervalo	Welch	20.48042	6.08831E-06	si
SQRT_acc_open_past_24mths	intervalo	Welch	95.9956	1.42722E-22	si
INV_both_mort_acc	intervalo	Welch	70.26624	5.81885E-17	si
SQRT_loan_amnt	intervalo	Welch	54.61581	1.57285E-13	si
INV_num_rev_tl_bal_gt_0	intervalo	Welch	20.13714	7.28315E-06	si
SQRT_open_rv_12m	intervalo	Welch	49.88526	1.72978E-12	si
EXP_pct_all_sats	intervalo	Welch	1.537188	0.2150637	no
SQR_pct_tl_nvr_dlq	intervalo	Welch	6.430033	0.01123452	si*
INV_pub_rec_bankruptcies	intervalo	Welch	5.588538	0.01809586	si*
INV_total_acc	intervalo	Welch	31.47035	2.07522E-08	si

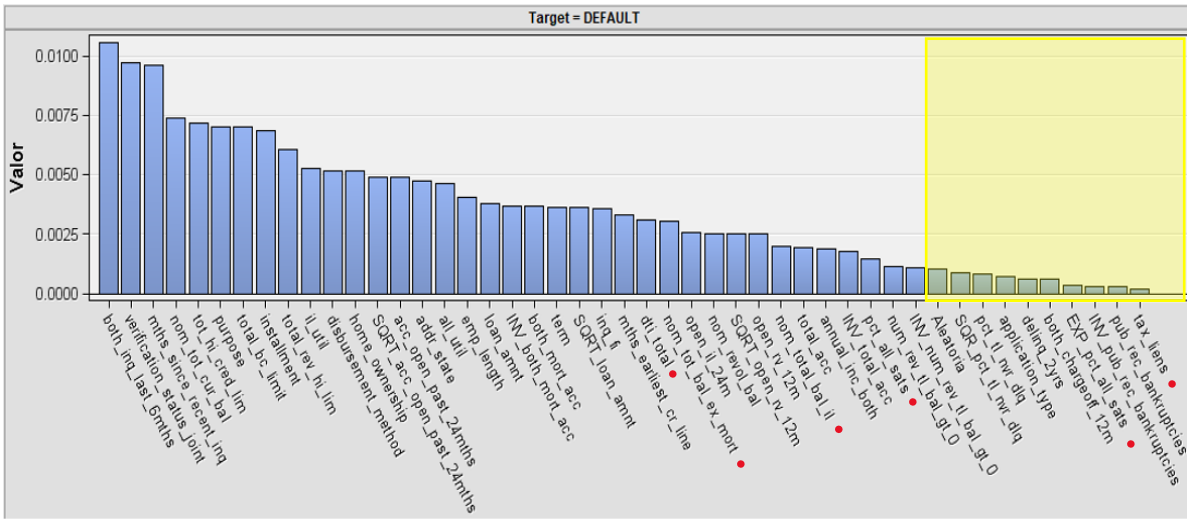
A continuación realizaremos otras pruebas de evaluación de aporte basadas en otros criterios.

[2] Las variables con especificación si* las mantendremos, pero están en el límite e incluso si consideráramos un nivel de significación más restrictivo quedarían fuera.

4.2.3 *Relación de las variables dependientes con la variable objetivo*

Para complementar los test anteriores utilizamos el nodo explorador de estadísticos en SAS Miner para ver la relación de las variables continuas y categóricas con la variable objetivo. De aquí podemos extraer otra clasificación de aporte de las variables basada en el estadístico Valor como se aprecia en el siguiente gráfico:

Figura 12: Gráfico Valor Importancia Variables



Hemos marcado con un punto rojo las variables que hemos identificado como no significativas con los test anteriores. Como podemos observar, ninguna de estas variables ocupa una posición en la cumbre del ranking de aporte pero no necesariamente las hubiéramos descartado todas basándonos en este estadístico. Además, creamos una variable aleatoria para ver cuáles son las variables que presentan menor importancia que esta y que, por lo tanto, presentarán aporte nulo al modelo (región amarilla). Por ello, utilizando este criterio nos desprenderíamos de las siete variables que quedan por debajo de la aleatoria.

Si nos fijamos en las variables con mayor aporte, situadas en la parte izquierda del gráfico vemos que coinciden con las que tenían mayor valor en los test del apartado anterior, lo cual permite deducir que estarán incluidas en la modelización y que serán la base de la composición de cualquier modelo. Las variables son: ***both_inq_last_6mths, verification_status_joint, mths_since_recent_inq, nom_tot_cur_bal, tot_hi_cred_lim, purpose, total_bc_limit, installment*** y ***total_rev_hi_lim***.

4.3 Agrupación de las variables: Credit Scoring

Cuando se crea un modelo de *credit scoring* se suele optar por la tramificación de las variables continuas para poder ver mejor la relación con la variable objetivo y obtener grupos sencillos que permitan clasificar el comportamiento de los individuos (Saddiqi, 2005).

Con esta técnica nos beneficiamos de la ganancia en interpretación y diferenciación de las características de los individuos para discriminar la mayor probabilidad de impago a partir de los tramos en los que se sitúa dicho individuo en cada una de las variables. Además, de esta forma las relaciones no lineales pueden captarse por modelos lineales. Este procedimiento se basa en asignar a cada una de las variables y a cada uno de sus tramos

un valor que refleja su aporte como predictor de comportamiento, conocido como WOE (*weight of evidence measure*). Esta medida refleja la diferencia entre la proporción de impago (1) y no impago (0) en cada atributo y mide la fuerza de cada grupo para separar los dos comportamientos. Cuanto mayor sea la diferencia en el WOE de grupos contiguos, mayor es la capacidad predictiva de ese atributo. Se calcula de la siguiente forma:

$$WOE_{attribute} = \ln \frac{p_{attribute}^{nonevent}}{p_{attribute}^{event}}$$

También es interesante ver con esta transformación como varía el poder predictivo de las variables. Para ello podemos usar los siguientes criterios:

- IV (*information value*): estadístico que ayuda a determinar el número de tramos y a estimar si una variable es buena para discriminar. Cuanto mayor sea este indicador mejor es la variable, teniendo en cuenta los siguientes criterios:
  - IV ≤ 0.02 →Variable no predictiva.
  - 0.02 ≤ IV ≤ 0.1 →Variable con poder predictivo débil.
  - 0.1≤ IV ≤ 0.3 →Variable con poder predictivo medio.
  - 0.3≤ IV ≤ 0.5→ Variable con poder predictivo fuerte.
  - 0.5 ≤ IV →Variable anómalamente predictiva.

  La forma de calcular este estadístico es la siguiente:

$$Information\ Value = \sum_{i=1}^{m} \left(P(attribute_i \mid non-event) - P(attribute_i\ event)\right) * WOE$$

- Índice de Gini: estadístico se usa también para verificar el poder discriminante de la variable. Toma valores entre 0 y 100%. Para los modelos de Credit Scoring se acota de la siguiente forma:
  - Gini ≤ 5% →variable no predictiva (descartada).
  - 5% ≤ Gini ≤ 15% →variable con poder predictivo medio.
  - 15% ≤ Gini →variable con poder predictivo alto.

  Este estadístico se calcula de la siguiente forma:

$$Gini\ Index = \left(1 - \frac{2 * \sum_{i=2}^{m} \left(n_i^{event} * \sum_{j=1}^{i-1} n_j^{non-event} \right) + \sum_{k=1}^{m} \left(n_k^{event} * n_k^{non-event} \right)}{N^{event} * N^{non-event}} \right) * 100$$

Nosotros nos basaremos en el criterio *information value* para determinar la entrada de variables por su posible aporte al modelo y establecemos un valor mínimo de entrada del 0,02. Por cuestión de espacio incluimos en el anexo IV la tramificación de las variables continuas y, si se ha realizado alguna recategorización para las categóricas, el WOE asignado a cada uno de los tramos identificados. De esta forma todas las variables serán tratadas en la modelización como variables continuas.

Como ya hemos mencionado, esta etapa del proceso también sirve como filtro de variables ya que se descartan un total de 20 por no cumplir con el criterio mínimo de entrada

estipulado. Para asegurarnos de que no estamos descartando variables que antes habíamos indicado ser buenas comprobamos su aporte en los test y los estadísticos analizados en los apartados anteriores. De esta forma verificamos que todas están entre aquellas que los test y la variable aleatoria descartaban, lo cual nos lleva a concluir que no se trata de variables con elevado poder predictivo y que podemos prescindir de ellas.

Figura 13: Variables Rechazadas en la Tramificación (criterio IV)

Variable	Nuevo Rol	Gini	IV
open_rv_12m	Rechazada	7.22	0.02
SQRT_open_rv_12m	Rechazada	7.29	0.02
nom_revol_bal	Rechazada	6.44	0.02
nom_tot_bal_ex_mort	Rechazada	5.15	0.02
total_acc	Rechazada	6.47	0.01
annual_inc_both	Rechazada	5.84	0.01
INV_total_acc	Rechazada	5.15	0.01
num_rev_tl_bal_gt_0	Rechazada	4.34	0.01
INV_num_rev_tl_bal_gt_0	Rechazada	4.37	0.01
pct_all_sats	Rechazada	4.41	0.01
EXP_pct_all_sats	Rechazada	4.68	0.01
nom_total_bal_il	Rechazada	4.41	0.01
application_type	Rechazada	2.51	0.01
pct_tl_nvr_dlq	Rechazada	3.05	0
SQR_pct_tl_nvr_dlq	Rechazada	2.83	0
pub_rec_bankruptcies	Rechazada	1.52	0
INV_pub_rec_bankruptcies	Rechazada	1.54	0
delinq_2yrs	Rechazada	1.34	0
tax_liens	Rechazada	0	0
both_chargeoff_12m	Rechazada	0	0

En este punto contamos con una muestra de los datos iniciales que contiene la información de 10.902 registros, 26 variables input y con la variable objetivo con el 50% de casos impago (1) y el 50% sin ningún impago (2).

4.4 Filtrado de variables para los algoritmos

A pesar de contar con un set reducido a 26 variables, consideramos oportuno conseguir un set de modelización más pequeño.

A diferencia de la regresión, la cual cuenta con su propio criterio de selección de variables, el resto de algoritmos que probamos utilizan todas las variables input que introduzcamos, por lo que podríamos estar creando modelos con demasiadas variables o con variables que no procede considerar. Por ello recurrimos a otros métodos de selección de variables que desvelan la importancia de las variables en cada uno de ellos y a partir de ello realizaremos un ranking general para quedarnos con aquellas que consideremos que aparentemente generarán aporte.

Para ello usamos diferentes métodos o modelos de prueba que harán de selectores dada la clasificación interna obtenida para las variables en cada uno de ellos. En SAS Miner utilizamos el nodo selección de variables, un árbol de decisión y un modelo Gradient Boosting. Además, utilizamos como criterio la mejor regresión obtenida en SAS Base utilizando validación cruzada repetida probando con distintos criterios de selección.

4.4.1 Mejor Regresión Logística

En este apartado llevamos a cabo la mejor regresión logística que podríamos generar para este set de datos actual con un total de 26 variables input. Como ya se ha comentado anteriormente, la regresión logística requiere hacer un procedimiento de descubrimiento iterativo para determinar cuáles son las variables más explicativas del modelo. Para ello se van a modelizar varias regresiones logísticas por varios métodos.

Antes de comenzar con la modelización, para facilitar el procedimiento y hacerlo más sencillo renombramos las variables tal y como presentamos en la siguiente figura:

Figura 14: Renombramos Variables para Modelizar

Nombre Original	Nuevo Nombre	Nombre Original	Nuevo Nombre
DEFAULT	Y	WOE_nom_tot_cur_bal	X14
WOE_INV_both_mort_acc	X1	WOE_open_il_24m	X15
WOE_SQRT_acc_open_past_24mths	X2	WOE_tot_hi_cred_lim	X16
WOE_SQRT_loan_amnt	X3	WOE_total_bc_limit	X17
WOE_acc_open_past_24mths	X4	WOE_total_rev_hi_lim	X18
WOE_all_util	X5	WOE_addr_state	X19
WOE_both_inq_last_6mths	X6	WOE_disbursement_method	X20
WOE_both_mort_acc	X7	WOE_emp_length	X21
WOE_dti_total	X8	WOE_home_ownership	X22
WOE_il_util	X9	WOE_mths_since_recent_inq	X23
WOE_inq_fi	X10	WOE_purpose	X24
WOE_installment	X11	WOE_term	X25
WOE_loan_amnt	X12	WOE_verification_status_joint	X26
WOE_mths_earliest_cr_line	X13		

Comenzamos entonces realizando una regresión con los tres criterios de entrada de variables *forward, backward* y *stepwise*, sin repetición. Con los tres obtenemos el mismo modelo con 17 variables que presentamos en la figura 15, lo cual a priori sin realizar validación cruzada ni repetición training test lleva a pensar que existe cierta estabilidad para el modelo que incluye este conjunto de datos. A pesar de ello, lo comprobaremos utilizando la validación cruzada repetida cuando tengamos otros modelos tentativos.

Figura 15: Mejor Modelo Regresión Logística Forward, Backward, Stepwise sin repetición

Modelo	Método selección	Repetición
X4 X5 X6 X8 X10 X11 X13 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26	Forward	1 sola vez
	Backward	
	Stepwise	

A continuación, utilizamos la macro %***randomselect*** (Portela, 2019) para tratar de encontrar los modelos que más se repiten utilizando distintos sets de datos *train* con el criterio de entrada de variables Stepwise. En este caso lo realizamos con 500 semillas diferentes y aquellos modelos que obtengamos con mayor frecuencia en todas estas pruebas, nos permitirá concluir cierta estabilidad y con más seguridad identificar que se trata de posibles buenos candidato a mejor modelo de regresión logística. El resultado es el siguiente:

Figura 16: Modelos Logística Stepwise %randomselectlog Mayor Frecuencia

	efecto	Frequency Count	Percent of Total Frequency
1	X4 X5 X6 X8 X11 X13 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26	148	29.540918164
2	X4 X5 X6 X8 X10 X11 X13 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26	109	21.756487026
3	X2 X5 X6 X8 X11 X13 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26	34	6.7864271457

El segundo modelo que más veces se repite coincide con el que hemos obtenido anteriormente con una sola repetición sin evaluar datos test, lo cual vuelve a aludir cierta estabilidad de este modelo. El que más veces se repite contiene una variable menos, X10 y el tercer modelo más frecuente se parece mucho al primero pero cambia la variable X4 por la X2, por lo que habrá que evaluarlos con validación cruzada repetida junto a otros posibles candidatos.

Finalmente, para tener otras alternativas y ver si pudiéramos igualar en sesgo y varianza con otros de dimensión más reducida buscamos el mejor modelo probando desde 8 a 17 variables. El resultado son diferentes modelos que cada vez van añadiendo una variable más, lo cual nos lleva a pensar que las variables base siempre serán las mismas y que se van introduciendo siguiendo un aporte creciente al modelo. Además, extraemos que los mejores modelos con 16 y 17 variables coinciden con los dos modelos más repetidos determinados en el paso previo. En la siguiente tabla se adjunta la composición de los modelos obtenidos con esta prueba:

Figura 17: Mejores modelos de 8 a 17 variables

Variables	Chi-cuadrado	Modelo
8	1027.77	X4 X6 X11 X16 X17 X19 X24 X26
9	1082.04	X4 X6 X11 X16 X17 X19 X21 X24 X26
10	1133.05	X4 X6 X11 X16 X17 X19 X21 X24 X25 X26
11	1172.44	X4 X6 X8 X11 X16 X17 X19 X21 X24 X25 X26
12	1212.57	X4 X5 X6 X8 X11 X16 X17 X19 X21 X24 X25 X26
13	1242.68	X4 X5 X6 X8 X11 X16 X17 X19 X20 X21 X24 X25 X26
14	1269.94	X4 X5 X6 X8 X11 X16 X17 X19 X20 X21 X23 X24 X25 X26
15	1282.27	X4 X5 X6 X8 X11 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26
16	1288.87	X4 X5 X6 X8 X11 X13 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26
17	1292.96	X4 X5 X6 X8 X10 X11 X13 X16 X17 X19 X20 X21 X22 X23 X24 X25 X26

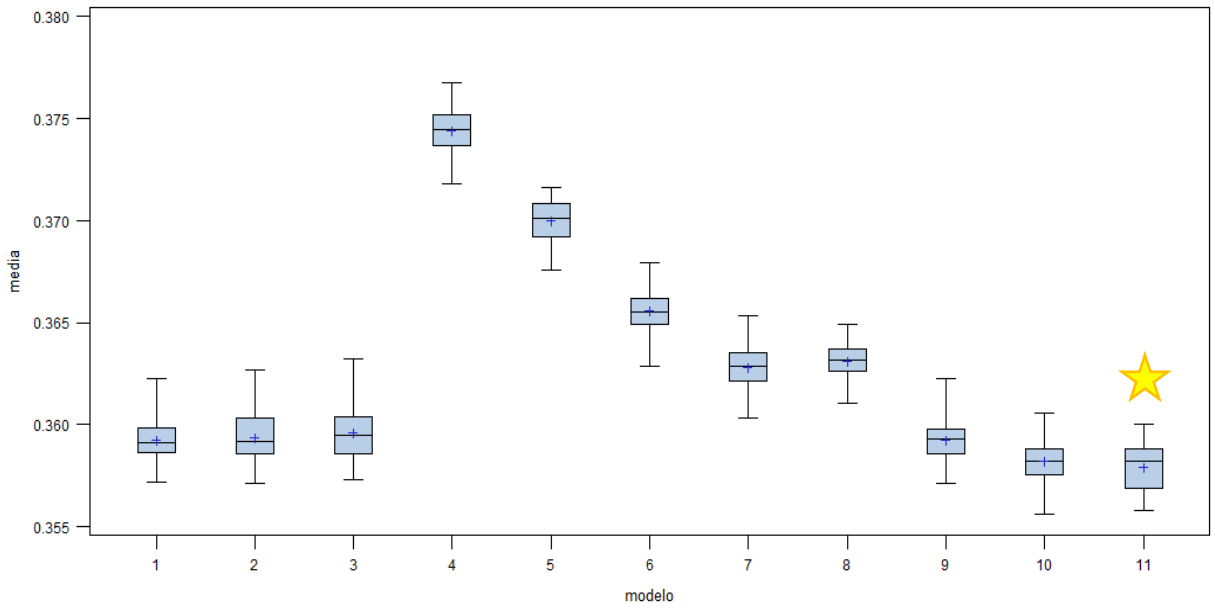
A priori ya podemos observar que quizás entre 8 y 13 variables son demasiado pocas por el nivel de estadístico chi-cuadrado asociado al modelo, pero quizás con 14 o 15 puede igualar e incluso mejorar los modelos con más parámetros. Lo comprobaremos a continuación con la macro %***cruzadalogística*** (Portela, 2019) que permite realizar la evaluación de modelos con validación cruzada para 4 grupos y 50 semillas. De esta forma nos aseguramos estabilidad de los resultados y podemos seleccionar no solo el menor nivel de sesgo, sino que también con un nivel apropiado de varianza. De esta forma conseguimos paliar el sesgo que puede generar la aleatoriedad en la selección del mejor modelo. En la figura 18 se incluye la descripción y composición de cada modelo y en la figura 19 su representación en diagramas de caja.

Finalmente, de la figura 19 podemos ver que dada la relación sesgo-varianza determinamos que la mejor opción es el modelo 11, ya que con 15 variables consigue mejorar los resultados obtenidos con 16 o 17 variables. Esta regresión la utilizamos como un criterio de selección de variables más y determinar así finalmente el set óptimo con el que proceder con la modelización del trabajo.

Figura 18: Tabla Descripción Regresión Logística Validación Cruzada

Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7	Modelo 8	Modelo 9	Modelo10	Modelo11
Simple/2º+rep	1º+repe	3º+repe	Mejor 8	Mejor 9	Mejor 10	Mejor 11	Mejor 12	Mejor 13	Mejor 14	Mejor 15
X4	X4	X2	X4	X4	X4	X4	X4	X4	X4	X4
X5	X5	X4	X6	X6	X6	X6	X5	X5	X5	X5
X6	X6	X5	X11	X11	X11	X8	X6	X6	X6	X6
X8	X8	X6	X16	X16	X16	X11	X8	X8	X8	X8
X10	X11	X8	X17	X17	X17	X16	X11	X11	X11	X11
X11	X13	X11	X19	X19	X19	X17	X16	X16	X16	X16
X13	X16	X13	X24	X21	X21	X19	X17	X17	X17	X17
X16	X17	X16	X26	X24	X24	X21	X19	X19	X19	X19
X17	X19	X17		X26	X25	X24	X21	X20	X20	X20
X19	X20	X19			X26	X25	X24	X21	X21	X21
X20	X21	X20				X26	X25	X24	X23	X22
X21	X22	X21					X26	X25	X24	X23
X22	X23	X22						X26	X25	X24
X23	X24	X23							X26	X25
X24	X25	X24								X26
X25	X26	X25								
X26		X26								

Figura 19: Box Plot Regresión Logística validación cruzada repetida



4.4.2 Ranking de filtrado

A continuación recopilamos en la figura 20 un resumen los resultados obtenidos con diferentes selectores previamente especificados que utilizamos en SAS Miner junto al resultado obtenido con la mejor regresión logística.

En este ranking de uso de las variables obtenemos un total de 13 variables que aparecen por consenso en todos los selectores utilizados, lo cual nos lleva a concluir que estas variables establecen el punto de partida y que debemos contar con ellas sin duda alguba, sea cual sea el modelo que estemos planteando. A continuación comprobamos para las variables que aparecen dos y tres veces su participación y aporte a cada uno de los criterios

de selección. Aquellas con aporte mínimo no las utilizaremos para evitar la sobre parametrización absurda, pero cuando se observe algún indicio de aporte mayor las mantendremos y evitaremos omitir información relevante:

- De las variables que aparecen en tres de los cuatro métodos mantenemos ***SQRT_loan_amnt*** (X3), ***nom_tot_cur_bal*** (X14), ***acc_open_past_24mths*** (X4) y ***il_util*** (X9) tras comprobar su posición de aporte en los distintos métodos. Por otra parte, comprobamos que es mejor prescindir del resto (marcamos en rojo en la tabla).
- De las variables que aparecen en dos de los cuatro selectores solamente mantenemos ***total_rev_hi_lim*** (X18) ya que presenta buen aporte en los dos métodos de árboles, los únicos que la incluyen en su proceso de modelización. Esto conlleva a deducir que en ella existe información no lineal que los otros dos métodos no captan pero que para otro tipo de algoritmos puede ser útil.
- La variable ***SQRT_acc_open_past_24mths*** (X2), solamente utilizada por el selector Gradient Boosting y tras comprobar su importancia en el modelo concluimos que no procede mantenerla para la modelización.

Figura 20: Selección de Variables para Algoritmos de Machine Learning

Variable		Nodo Selección*	Mejor Regresión	Árbol	GB	RANK
all_util	X5	1	1	1	1	**4**
both_inq_last_6mths	X6	1	1	1	1	**4**
dti_total	X8	1	1	1	1	**4**
installment	X11	1	1	1	1	**4**
tot_hi_cred_lim	X16	1	1	1	1	**4**
total_bc_limit	X17	1	1	1	1	**4**
addr_state	X19	1	1	1	1	**4**
emp_length	X21	1	1	1	1	**4**
home_ownership	X22	1	1	1	1	**4**
mths_since_recent_inq	X23	1	1	1	1	**4**
purpose	X24	1	1	1	1	**4**
term	X25	1	1	1	1	**4**
verification_status_joint	X26	1	1	1	1	**4**
INV_both_mort_acc	X1	1	0	1	1	**3**
SQRT_loan_amnt	X3	1	0	1	1	**3**
mths_earliest_cr_line	X13	1	0	1	1	**3**
nom_tot_cur_bal	X14	1	0	1	1	**3**
acc_open_past_24mths	X4	1	1	0	1	**3**
il_util	X9	1	0	1	1	**3**
open_il_24m	X15	1	0	1	1	**3**
disbursement_method	X20	1	1	1	0	**3**
loan_amnt	X12	1	0	0	1	**2**
both_mort_acc	X7	1	0	0	1	**2**
inq_fi	X10	0	0	1	1	**2**
total_rev_hi_lim	X18	0	0	1	1	**2**
SQRT_acc_open_past_24mths	X2	0	0	0	1	**1**

**El criterio utilizado para determinar la entrada o salida con el nodo de selección de variables es el de R cuadrado*

En la tabla vemos las variables que compondrán el set de modelización definitivo (marcadas en negro). Este set es con el que evaluaremos todos y cada uno de los diferentes algoritmos que queremos considerar, introduciendo las mismas 18 variables input para todos para partir de igualdad de condiciones en el procedimiento.

5 Modelización y comparación de modelos

Para la modelización se va a utilizar la regresión logística y los algoritmos de *Machine Learning* previamente introducidos: redes neuronales, *bagging, random forest, gradient boosting* y *SVM*. Además, realizaremos pruebas de ensamblado para tratar de mejorar los mejores modelos obtenidos con cada algoritmo y compararemos con validación cruzada repetida para llegar a conclusiones válidas. A partir de ahora, para todos los procedimientos se utilizará la validación cruzada repetida para cuatro grupos y repetición con 20 semillas.

5.1 Red Neuronal

Tal y como hemos mencionado en la explicación teórica de este algoritmo se trata de una opción potente y robusta pero requiere una buena parametrización y muchas pruebas para conseguir alcanzar un buen funcionamiento. A priori, no sabemos si este algoritmo conseguirá hacerle sombra al modelo clásico utilizado por excelencia para los modelos de *credit scoring*, por lo que trataremos de realizar buenas configuraciones que nos permita superarla o al menos acercarnos a ella.

El primer parámetro que configuramos es el número de nodos, el cual tiene elevada dependencia de la caracterización del set de datos, especialmente de su complejidad. Como trabajaremos con 18 variables tendremos que adaptar el número de nodos basándonos en el siguiente criterio: establecemos que, como mínimo, deberán mantenerse entre 20 y 30 observaciones por parámetro de la red, por lo que tendremos que coger el total de observaciones del conjunto de datos entrenamiento (70%) y calcular cuántos nodos es óptimo aplicar con la fórmula que se presenta a continuación. Contamos entonces con unas 7.631 observaciones para el entrenamiento y las 2.701 restantes servirán para la validación:

$$Nº\ parámetros = h\ (k+1)+h+1$$

Donde h = nodos ocultos, k = nodos input (parámetros que introducimos) y estableciendo el criterio de que queremos mantener mínimo 30 observaciones por parámetro y tenemos 7.631, el número de parámetros adecuado estará en torno a $\frac{7631}{30} \approx 254$.

Entonces con 18 variables o nodos input = k y el número de parámetros= 254:

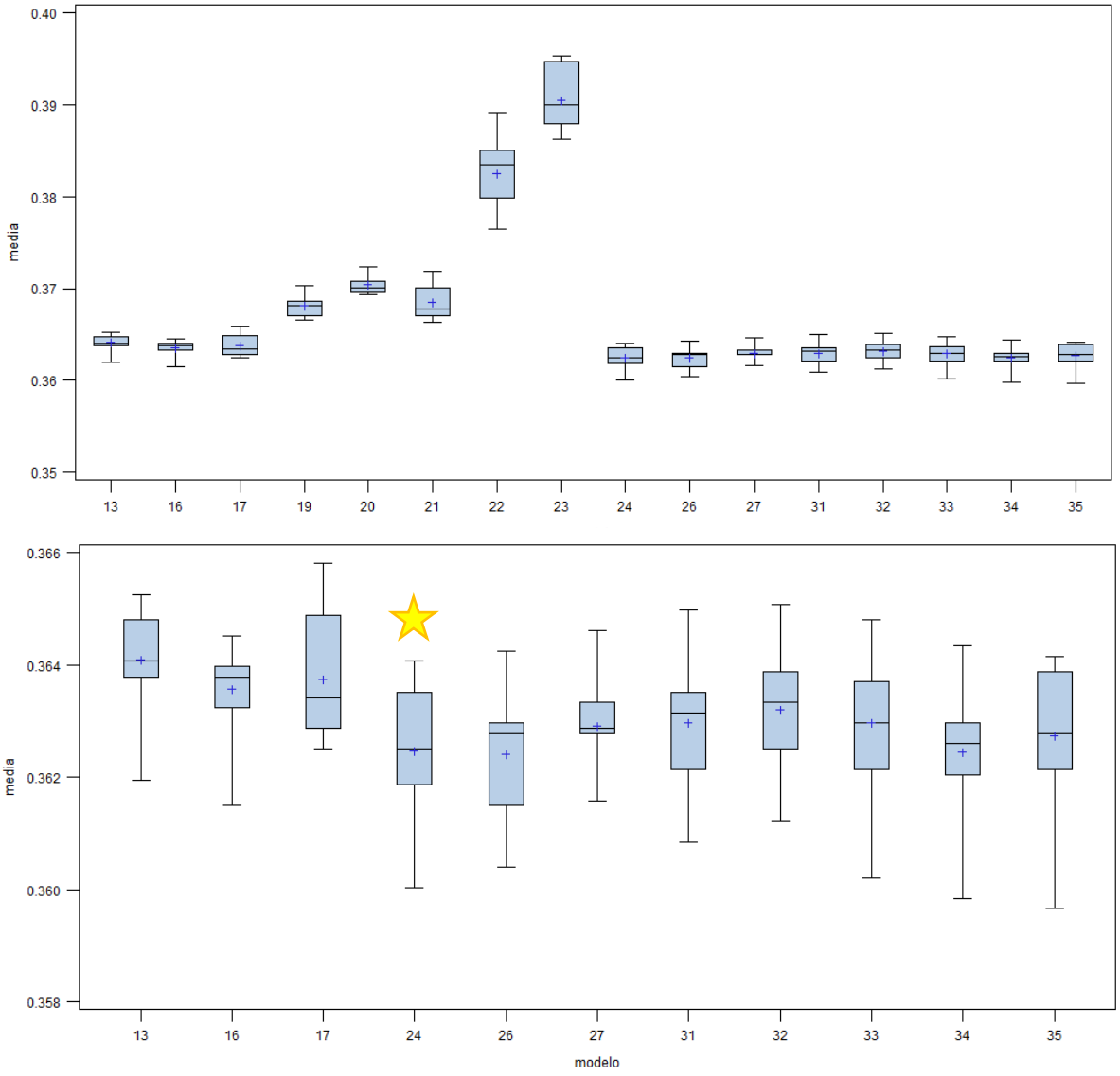
$$254 = h\ (18+1)+h+1 \rightarrow h \approx 12/13\ nodos\ ocultos$$

Tras realizar algunas exploraciones con diferente número de nodos, podemos confirmar que el óptimo se mueve en torno a 12/13/14 pero para asegurarnos realizaremos pruebas con validación cruzada repetida con diferentes valores para este parámetro. Para ello utilizamos la macro %***cruzadasbinarianeural*** (Portela, 2019) y realizaremos pruebas con otras configuraciones como por ejemplo utilizando redes con más nodos pero estableciendo *early stopping* (parada anticipada de sus iteraciones). Probaremos también con diferentes algoritmos de optimización y funciones de activación. A continuación adjuntamos una tabla resumen (figura 21) con las diferentes pruebas realizadas y su representación utilizando el diagrama de caja (figura 22) con el objetivo escoger la mejor configuración para los datos.

Figura 21: Tabla Configuración Redes Neuronales

Modelo	Nodos	Algoritmo	Momentum	Learn. Rate	F. activ.	Early stop.
13	11	bprop	0.2	0.1	tanh	.
16	14	bprop	0.2	0.1	tanh	.
17	15	bprop	0.2	0.1	tanh	.
19	18	bprop	0.2	0.1	tanh	60
20	19	bprop	0.2	0.1	tanh	50
21	20	bprop	0.2	0.1	tanh	40
22	11	quanew	.	.	tanh	.
23	14	quanew	.	.	tanh	.
24	14	bprop	0.8	0.2	tanh	.
26	10	bprop	0.8	0.2	tanh	.
27	11	bprop	0.8	0.2	tanh	.
31	16	bprop	0.8	0.2	tanh	.
32	14	bprop	0.8	0.2	log	.
33	14	bprop	0.8	0.2	lin	.
34	14	bprop	0.8	0.2	arc	.
35	14	bprop	0.8	0.2	sin	.

Figura 22: Diagrama de Cajas Redes Neuronales



El segundo diagrama de cajas es una ampliación del primero excluyendo aquellas redes con un error medio muy por encima del resto que no permiten evaluar y seleccionar la mejor red neuronal. A pesar realizar muchas pruebas, cambiando mucho los parámetros, observamos como los resultados no varían de forma significativa pero sí es cierto que si nos centramos en elegir la red con mejor relación nivel de sesgo y varianza escogemos la red neuronal número 24. Ésta contiene 14 nodos ocultos, algoritmo *backpropagation* con *momentum* 0.8 y *learning rate* 0.2 y algoritmo de optimización tangente hiperbólica; su error medio es de 0,36236. Cabe mencionar que hemos realizado pruebas de *early stopping* para ver si mejorábamos el nivel de varianza pero era contraproducente el deterioro producido en el sesgo, por lo que finalmente deducimos que no procede establecer criterio de parada.

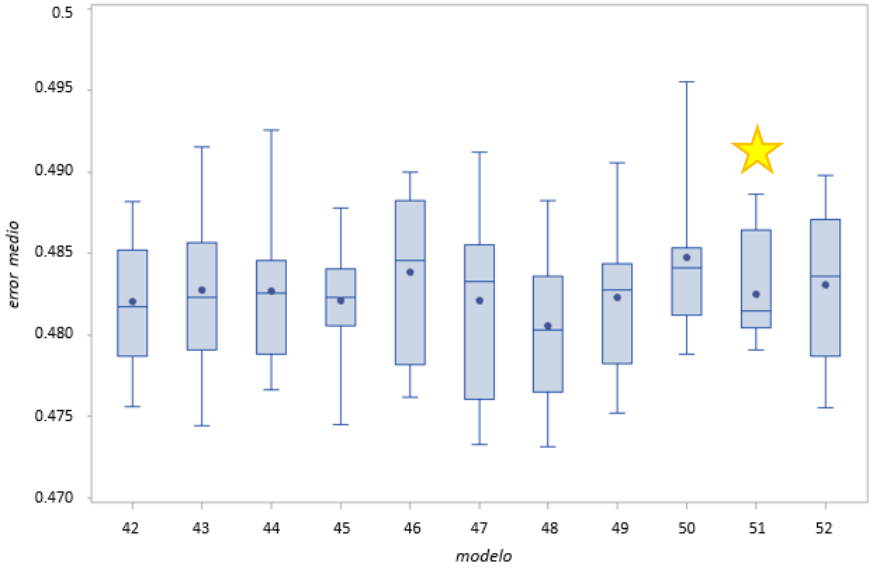
5.2 Bagging

Para el modelo basado en árboles *bagging* utilizamos la macro de SAS %***cruzadarandomforestbin*** (Portela, 2019). En esta macro se da la posibilidad de probar y modificar los siguientes parámetros: variar el porcentaje total de la muestra utilizado para construir los árboles, el tamaño de observaciones mínimo estipulado para las hojas finales, la profundidad máxima del árbol y el p-valor establecido para abrir nuevos nodos. Tras realizar diferentes pruebas y ajustes, estas son las configuraciones principales con sus respectivas representaciones gráficas:

Figura 23: Tabla Configuración Bagging

Modelo	Variables	% Muestra	Tamaño hoja	Prof. Max.	p-valor
41	18	0.75	25	10	0.1
42	18	0.75	15	20	0.1
43	18	0.75	35	10	0.1
44	18	0.5	25	10	0.1
45	18	0.25	25	10	0.1
46	18	0.9	25	10	0.1
47	18	0.75	25	10	0.05
48	18	0.75	15	20	0.01
49	18	0.75	25	10	0.2
50	18	0.75	40	6	0.1
51	18	0.5	30	15	0.1
52	18	0.75	40	15	0.1

Figura 24: Diagrama de Cajas Bagging



El mejor modelo de *bagging* es el 51 por ser el que presenta una mejor combinación de sesgo y varianza. Está creado con un 50% de porcentaje de muestra, con un tamaño mínimo de hoja final de 30 observaciones, profundidad máxima de los árboles de 15 niveles y p-valor de apertura de nodo de 0,1. Cabe mencionar que el nivel medio de error aumenta significativamente respecto a los que alcanzábamos con la red y la regresión, situándonos en torno a 0,48 en el mejor caso. Tendremos que comprobar si con los otros algoritmos basados en árboles sucede lo mismo o si los modelos basados en árboles no ajustan bien estos datos.

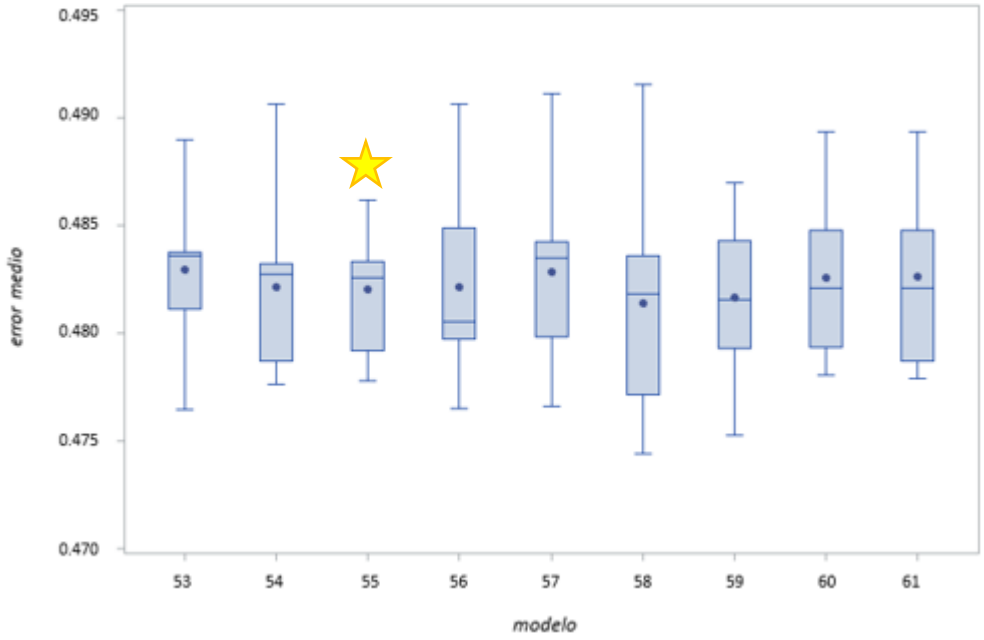
5.3 Random Forest

Random forest es una variante del procedimiento de modelización de *bagging*, el cual se caracteriza por sortear aleatoriamente un número de variables inferior al total de la muestra para la apertura de cada nodo. En el caso de *bagging* hemos visto que N=18, es decir, el total de variables de la muestra de modelización, mientras que en este caso probaremos distintas n < N y especificaremos la más adecuada junto a la parametrización del resto de características a determinar. Para este modelo se utiliza la misma macro de SAS que para *bagging* pero simplemente modificaremos el parámetro que hace referencia al número de variables a sortear en la apertura de cada nodo. Las configuraciones que hemos probado son las siguientes:

Figura 25: Tabla Configuración Random Forest

Modelo	Variables Sorteadas	Porcentaje Muestra	Tamaño Min. Hoja	Profundidad Máx.	P-valor
53	8	0.5	25	10	0.1
54	4	0.75	25	10	0.1
55	3	0.75	25	10	0.1
56	12	0.75	25	10	0.1
57	15	0.75	25	10	0.1
58	17	0.75	15	20	0.1
59	3	0.5	15	20	0.1
60	3	0.5	20	15	0.1
61	3	0.5	20	15	0.1

Figura 26: Diagrama de Cajas Random Forest



El mejor modelo de Random Forest se crea estableciendo que en la apertura de cada nodo se utilicen tres variables diferentes, con un porcentaje de la muestra del 75%, un tamaño mínimo de observaciones en últimos nodos de 25 observaciones, una profundidad máxima de 10 y un p-valor de 0,1. Como podemos observar el nivel de error medio sigue siendo elevado y se mantiene en torno a 0,40, por lo que por ahora los dos modelos basados en árboles probados no son acertados para llevar a cabo buenas predicciones para los datos.

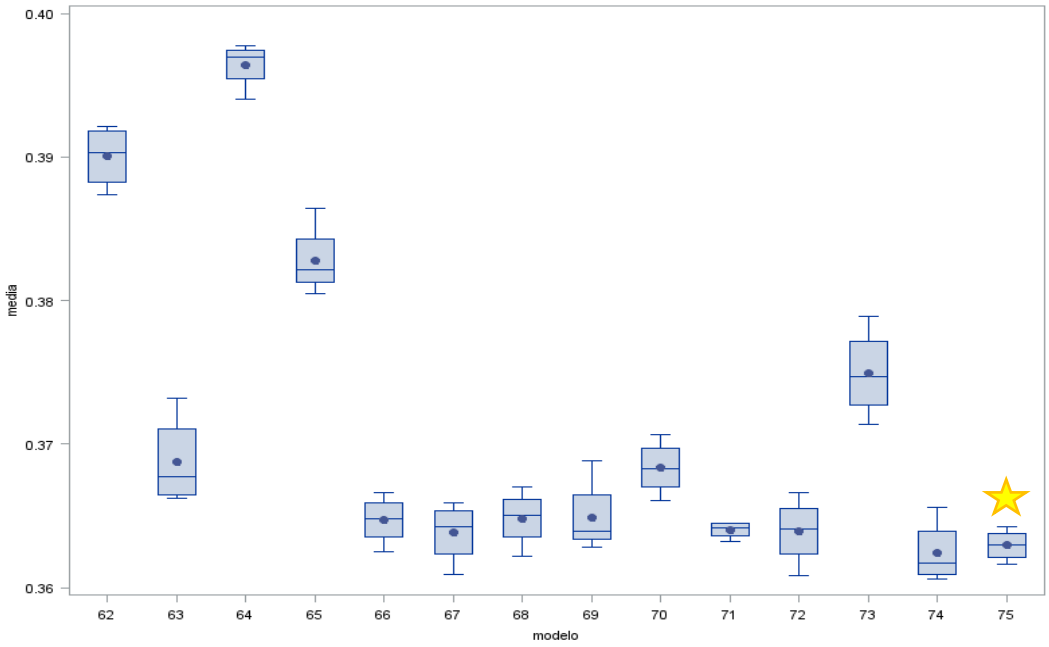
5.4 Gradient Boosting

Para este último algoritmo basado en árboles cabe tener en mente que se van construyendo iterativamente y poco a poco se va modificando la predicción inicial en cada paso con el objetivo de minimizar los residuos en sentido decreciente con tendencia a cero. En este algoritmo entonces se introduce un parámetro que gradúa cuánto se va corrigiendo el error en cada iteración, por lo que se trata de encontrar una combinación de número de iteraciones y esta tasa de aprendizaje adecuada. Cabe considerar que para tasas de aprendizaje muy pequeñas son necesarias más iteraciones, mientras que para tasas moderadas el número de iteraciones necesarias es inferior.

Figura 27: Tabla Configuración Random Forest

Modelo	Tasa Aprendizaje	Tamaño Hoja Final	Profundidad Máxima	Obs. Mín. Nuevo Nodo
62	0.05	10	20	20
63	0.01	10	20	20
64	0.1	10	20	20
65	0.03	10	20	20
66	0.01	20	20	20
67	0.01	50	20	30
68	0.01	50	10	30
69	0.01	30	20	30
70	0.01	10	15	30
71	0.01	40	20	30
72	0.01	50	15	20
73	0.001	50	10	30
74	0.015	50	10	30
75	0.02	50	10	30

Figura 28: Diagrama de Cajas Gradient Boosting



Para este algoritmo utilizamos la macro de SAS ***%cruzadarandomforestbin*** (Portela, 2019) y realizamos diferentes pruebas con todas las modificaciones que esta permite probar. Las anteriores tablas resumen las exploraciones realizadas para conseguir la parametrización óptima del algoritmo y su respectiva representación en un diagrama de cajas. Como podemos observar, la configuración con mejor relación entre sesgo y varianza presenta es el modelo 75, con una tasa de aprendizaje 0,02, un tamaño mínimo de hoja final de 50 observaciones, una profundidad máxima de árbol de 10 y un mínimo de observaciones para abrir un nuevo nodo igual a 30. Este algoritmo basado en arboles sí consigue alcanzar e incluso mejorar los niveles de error que obteníamos con la red neuronal, por lo que descartamos que los modelos basados en árboles no sean buenos para modelizar este set de datos y especificamos que Gradient Boosting podría ser un posible modelo ganador.

5.5 Support Vector Machines

Este caso se trata de un conjunto de algoritmos de aprendizaje supervisado en los que, dado un conjunto de puntos donde cada uno pertenece a una de las dos posibles clases objetivo, el algoritmo construye un modelo que separa los puntos nuevos (cuya clase desconocemos) en una categoría u otra. Como hemos visto, existen diferentes tipos de funciones *kernel* para llevar a cabo la separación pero para nuestros datos la única que converge y nos permite llevar a cabo la separación en SAS Base es *SVM* lineal, por lo que desarrollamos diferentes modelos variando el parámetro C que representa el inverso al margen de error que cedemos cometer en la clasificación de los puntos. Para desarrollar este algoritmo con el *kernel* lineal utilizamos la macro ***%cruzadasSVMbin*** (Portela, 2019) en la que podemos modificar el parámetro de control del margen de error. A continuación adjuntamos la configuración de los modelos probados con su representación de error medio:

Figura 29: Diagrama de Cajas SVM

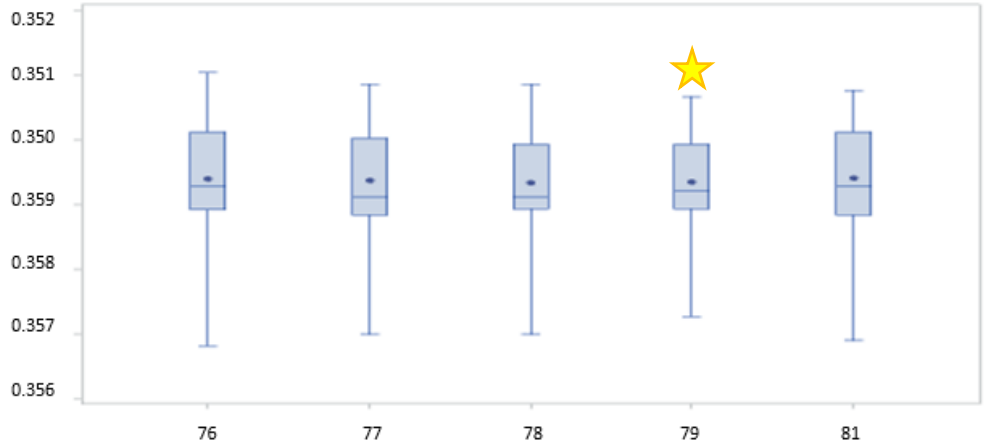


Figura 30: Tabla Configuración SVM

Modelo	*Kernel*	Inverso al Margen de Error
76	linear	50
77	linear	20
78	linear	8
79	linear	0.8
80	linear	0.2
81	linear	100

Finalmente, concluimos que la mejor opción es establecer un margen de error elevado mediante la parametrización de C en 0,08 ya que, a pesar de que todas las pruebas realizadas presentan niveles de error medios muy semejantes, con esta configuración se reduce algo la varianza. Además, observamos como este algoritmo alcanzamos un error medio de 0,3593, lo cual muestra ser un nivel de error competitivo con respecto al de los demás algoritmos desarrollados.

5.6 Regresión Logística

Finalmente, cabe mencionar que la mejor regresión logística obtenida al principio del estudio para utilizarla como un método de filtrado presenta una cierta ventaja a la hora de comparar este modelo con el resto de los algoritmos probados. En efecto, aquel modelo de regresión logística contiene la combinación de variables que mejor se adapta a su funcionamiento ya que utilizó un procedimiento de estimación por pasos, por lo que a efectos puramente comparativos y debido a que las diferencias entre los diferentes modelos son muy pequeñas, hemos decidido crear un modelo de regresión logística sin criterio de selección. De esta forma, al igual que hemos hecho con el resto de algoritmos, generamos un modelo que simplemente evalúe una regresión creada con el set de 18 variables introducidas como explicativas. Por ello creamos el modelo 82 que evaluaremos y compararemos con el resto utilizando validación cruzada repetida. Además, mantendremos la mejor regresión logística obtenida al principio del estudio para valorar la diferencia entre ambas. Como es lógico, la nueva regresión logística generada en este apartado no proporcionará un ajuste tan bueno como el inicial, que elige entre todas las variables del set inicial de partida aquellas que resultan significativas.

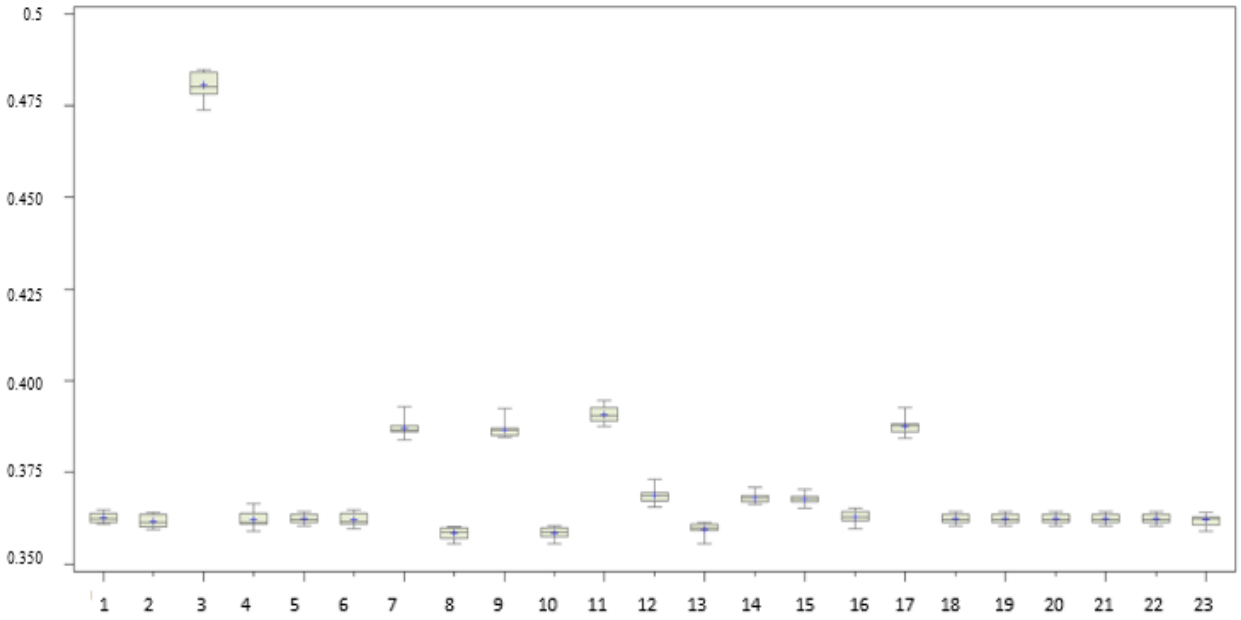
5.7 Ensamblado de modelos en SAS Base

En este apartado finalmente realizamos unas pruebas de ensamblado utilizando la configuración de los mejores modelos obtenidos en cada uno de los apartados anteriores. Se usa la macro ***%cruzadastack*** (Portela, 2019) la cual vuelve a calcular todos los modelos con la configuración que se desee introducir, usando el mismo set de variables para todos, mismas semillas y mismos grupos de validación cruzada. El ensamblado, además de mostrar y recalcular los modelos de forma individual, realiza combinaciones de los modelos para tratar mejorar el error. Por ello en este paso veremos si el cálculo de los algoritmos utilizando esta macro varía los resultados respecto a los cálculos individuales de cada uno y si compensa complicar aún más los modelos realizando combinaciones o si es preferible quedarnos con la predicción simple obtenida con un único modelo. A continuación adjuntamos una tabla resumen de las pruebas realizadas y su representación.

Figura 31: Tabla Resumen Modelos Ensamblado

Modelo	Ensamble	Modelo	Ensamble
1	Logistica	13	Logistica- Red- GBoosting
2	Red	14	Logistica- RForest- GBoosting
3	RForest	15	Red- RForest - GBoosting
4	GBoosting	16	Logistica- Red- RForest- GBoosting
5	SVM lin.	17	Logistica*0.2- Red*0.1- RForest*0.5- GBoosting*0.2
6	Logistica-Red	18	Logistica-SVM
7	Logistica-RForest	19	GBoosting-SVM
8	Logistica - GBoosting	20	Red-SVM
9	Red - RForest	21	GBoosting - SVM
10	Red - GBoosting	22	SVM- Red- RForest
11	RForest - GBoosting	23	Logistica-Red-RForest-GBoosting-SVM
12	Logistica - Red - RForest		

Figura 32: Diagrama de Cajas Ensamblado de Modelos

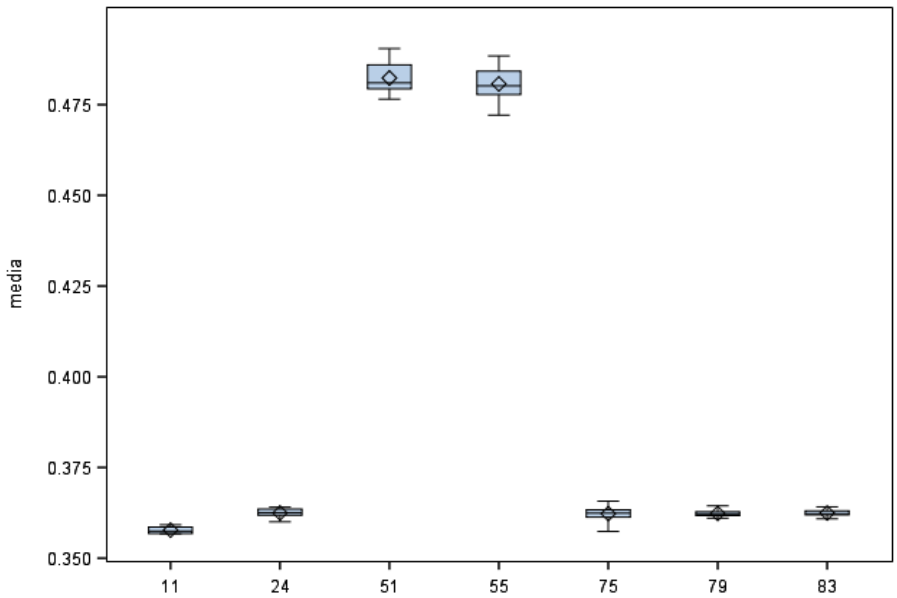


Los modelos 1, 2, 3, 4 y 5 son los algoritmos simples y observamos resultados similares en cuanto al nivel de error obtenido en los apartados anteriores. Como podemos observar una vez más, *random forest* es la peor opción y todos los ensamblados que lo incluyen empeoran algo el error cometido. Por otra parte, a pesar de que el ensamblado reduce el nivel de error medio algo, se trata de una mejora muy pequeña y concluimos que no compensa el aumento de complejidad que genera en los modelos.

5.8 Selección del Mejor Modelo en SAS

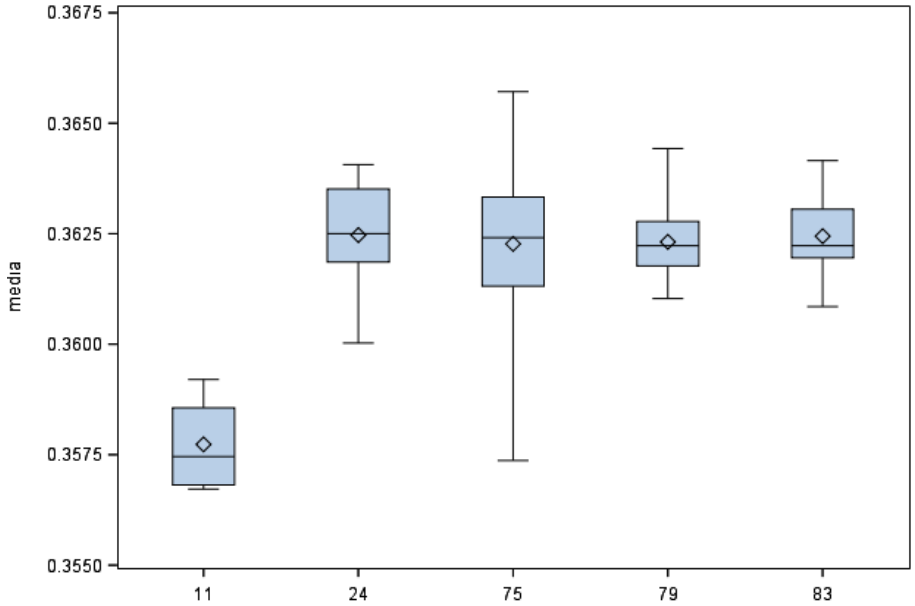
En este apartado procedemos a evaluar cuál es el mejor modelo para tratar este set de datos. Para ello ponemos a competir la mejor opción obtenida para cada algoritmo en validación cruzada repetida utilizando 100 semillas y 10 grupos. Como resultado obtenemos el siguiente diagrama de cajas:

Figura 33: Diagrama de Cajas Mejor Modelo SAS (1)



Los dos modelos de Bagging y Random Forest, tal y como ya habíamos identificado previamente, presentan niveles de error bastante más elevados en comparación al resto de los algoritmos, por lo que los descartamos para poder valorar mejor al resto y seguir con la tarea de escoger el mejor modelo dada la combinación de sesgo, varianza y complejidad:

Figura 34:Diagrama de Cajas Mejor Modelo SAS (2)



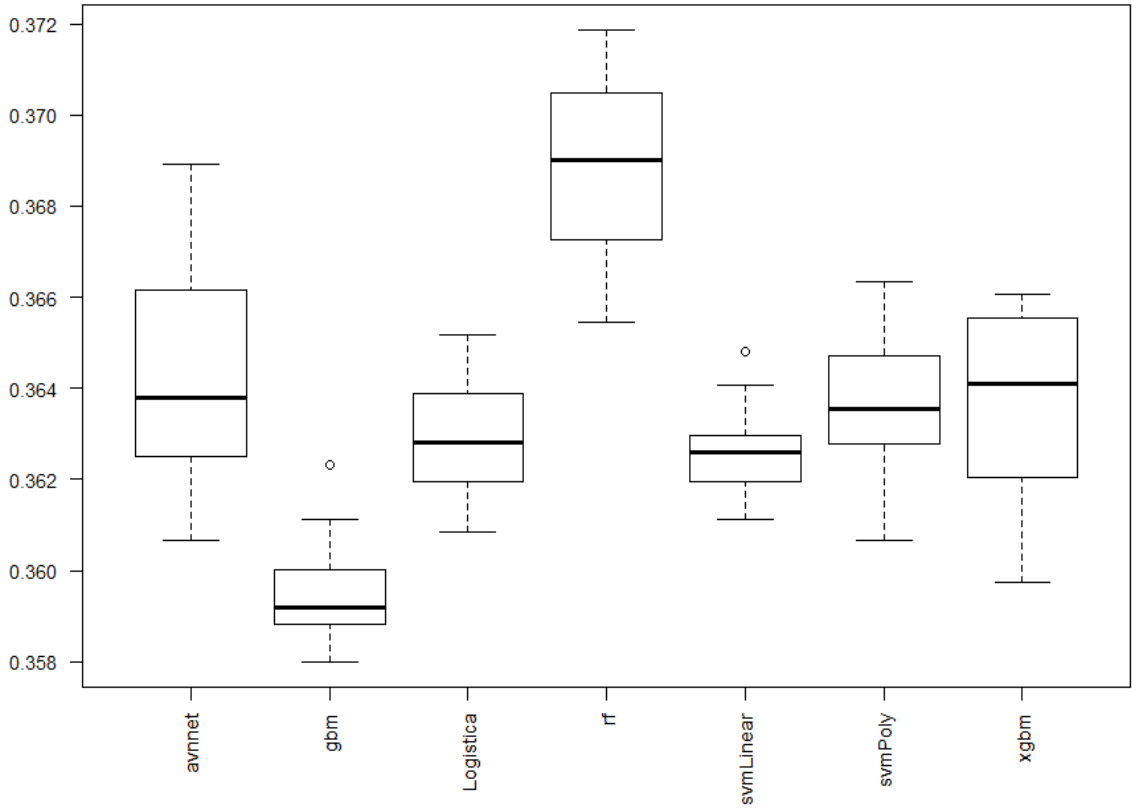
De este diagrama ampliado se extraen varias conclusiones:

1. Si no tenemos en cuenta el criterio de igualdad de condiciones en cuanto a las variables de entrada para realizar los modelos, la mejor opción es la regresión logística 11 (propio criterio de selección de variables). En realidad, a esta regresión le cedemos libertad para que sea ella misma la que escoja aquellas variables que le permitan optimizar el procedimiento, por lo que no es equitativo ni realista basarnos en este criterio para seleccionar la regresión como mejor alternativa para la modelización. Si hiciéramos lo mismo con cada algoritmo y adaptáramos a ellos las variables de entrada, seguramente consiguiéramos el mismo efecto en todos.
2. Por otra parte, si forzamos que la logística introduzca las mismas variables con las que estamos formulando el resto de los algoritmos (regresión 83) podemos observar como pierde la primera posición y se coloca en línea de tasa de fallo con el resto de los algoritmos evaluados: red (24), *gradient boosting* (75) y *SVM* (79).
3. Finalmente, tras evaluar en igualdad de condiciones los diferentes modelos vemos como los niveles de sesgo y varianza que ofrece la logística son competitivos y consiguen igualar e incluso superar a los algoritmos de *Machine Learning*. De aquí sí extraemos un criterio justo para determinamos que la mejor opción es utilizar este algoritmo clásico ya que no solo permite mantener cierta sencillez en la modelización, sino que también permite mantener la interpretabilidad de los resultados, lo cual es muy útil para este tipo de trabajos.

5.9 Desarrollo del estudio en R

Paralelamente hemos realizado un estudio similar con el programa R para evaluar cómo podrían variar los resultados tras realizar todo este procedimiento utilizando otro software. En R contamos con una serie de paquetes y funciones equivalentes a las macros utilizadas en SAS que requieren realizar el mismo proceso exhausto de configuración y pruebas para todos los algoritmos. Existen algunas diferencias de configuración que hacen que el desarrollo no sea exactamente igual, por lo que hemos tenido que repetir el procedimiento de prueba con diferentes posibles configuraciones[3] hasta encontrar la combinación óptima para cada algoritmo. Además, estas diferencias en la configuración permiten evaluar también qué programa permite alcanzar mejor relación sesgo-varianza. A continuación representaremos el diagrama de cajas final obtenido en R en el que comparamos la tasa de error del mejor modelo obtenido para cada algoritmo utilizando las mismas 18 variables. También adjuntamos una tabla resumen de las configuraciones que contiene cada uno de ellos.

Figura 35: *Tabla resumen Mejores Modelos R*



[3] En anexos se incluye el acceso al código utilizado con todas las pruebas y los resultados obtenidos para cada algoritmo.

Figura 36: Diagrama de Cajas Mejor Modelo R Tasa Fallo

MEJOR MODELO EN R TABLA RESUMEN				
Red Neuronal = avnnet				
	Nodos	**Tasa Aprendizaje**		
	8	0.1		
Random Forest= rf				
Árboles	**Tamaño Muestra**	**Obs. Mín. Nodo**	**Profundidad Max.**	**Variables Sortear**
500	1088	30	10	12
Gradient Boosting paquete gbm = gbm				
Árboles	**Tasa Aprendizaje**	**Obs. Mín. Nodo**	**Profundidad Max.**	
3000	0.03	25	10	
Gradient Boosting paquete XGBoost = xgbm				
Árboles	**Tasa Aprendizaje**	**Obs. Mín. Nodo**	**Profundidad Max.**	
100	0.05	50	10	
**Vemos que otros parámetros de regularización internos del algoritmo es mejor mantenerlos por defecto*				
SVM lineal = SVMlinear				
		Inverso Margen Error		
		0.05		
SVM Polinomial = SVMPoly				
	Inv. Margen Error	**Grado del Polinomio**	**Escala**	
	0.01	2	0.5	

De esta comparación de modelos podemos extraer varias conclusiones:

1. A simple vista, si nos fijamos en el nivel de error medio en torno al cual se sitúan los algoritmos vemos como este es muy parecido al alcanzado con SAS.
2. A diferencia de lo observado en la comparación de modelos de SAS, el mejor modelo nos lo genera el algoritmo *gradient boosting* configurado con el paquete de R gbm, con un error medio de 0,359. Esto nos lleva a concluir que el procedimiento de modelización de algoritmos basados en árboles en R, especialmente de *gradient boosting*, supera las posibilidades ofrecidas por SAS y da lugar a resultados con mejor relación sesgo-varianza.
3. Los siguientes modelos del ranking son la regresión logística y *SVM* con *kernel* lineal. Esta regresión la hemos configurado con los mismos parámetros que hemos formulado los algoritmos de *Machine Learning*, por lo que se encuentra en igualdad de condiciones y vemos como sigue ocupando posición en el pódium.
4. El peor modelo, una vez más, es *random forest*. Esto permite concluir que este algoritmo basado en árboles (al igual que la variante *bagging*) no consiguen adaptar bien nuestros datos y que en este caso no procede considerarlos como alternativa de modelización.
5. En este caso, la red neuronal configurada con *avnnet* de Caret no consigue igualar la relación sesgo-varianza alcanzada en SAS, por lo que concluimos que para este algoritmo se facilitan alternativas de configuración mejores en SAS que en R.
6. Cabe valorar una vez más si la mejora del error que supone el modelo de *gradient boosting* respecto a la logística y a la pérdida de interpretabilidad que esto supone. Para estos datos la diferencia es de menos del 0,2% de error, por lo que concluimos que no compensa.

A continuación realizamos también unas pruebas de ensamblado con estos mejores modelos en R para ver si conseguimos mejoras importantes mediante la configuración de esta técnica en R. El resultado es el siguiente:

Figura 37: Diagrama de Cajas Ensamblado R

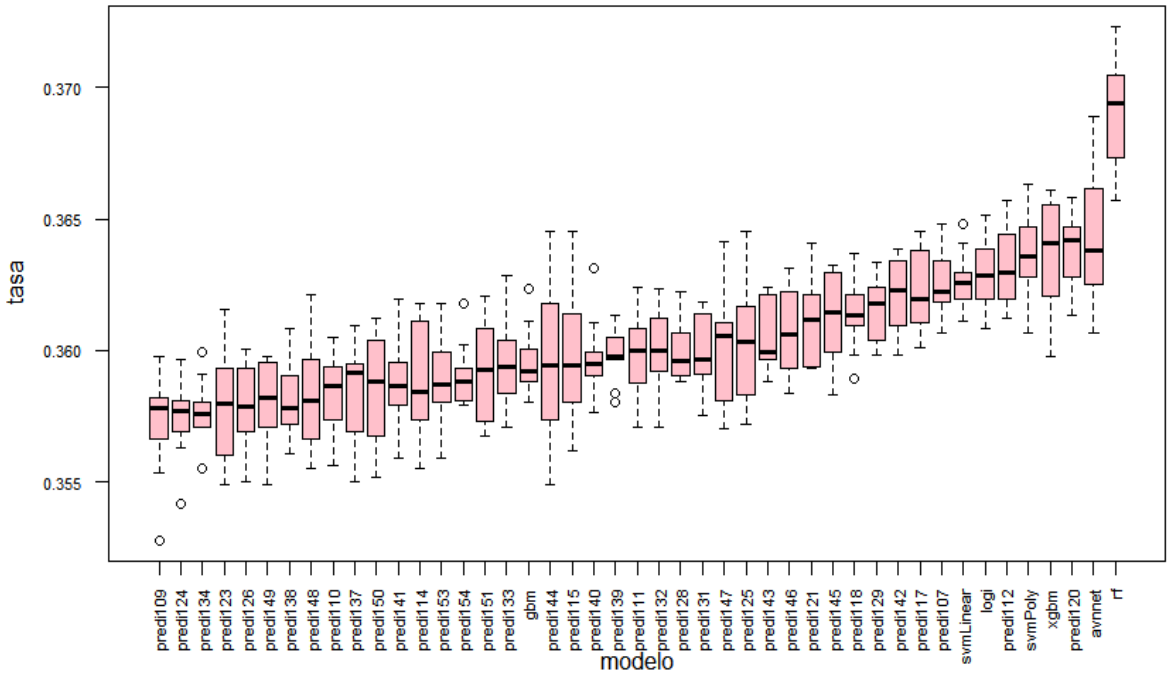


Figura 38: Composición Modelos Ensamblado R

Modelo	Componente1	Componente2	Modelo	Componente 1	Componente2	Componente3
logi	Logistica	.	predi125	Logistica	Red	RF
avnnet	Red	.	predi126	Logistica	Red	GB(gbm)
rf	Random Forest	.	predi127	Logistica	Red	GB(Xgboost)
gbm	Gradient Boosting (gbm)	.	predi128	Logistica	Red	SVMPoly
xgb	Gradient Boosting (Xgboost)	.	predi129	Logistica	Red	SVMLinear
SVMLinear	SVM Linear	.	predi130	Logistica	RF	GB(gbm)
SVMPoly	SVM Polynomial	.	predi131	Logistica	RF	GB(Xgboost)
predi107	Logistica	Red	predi132	Logistica	RF	SVMPoly
predi108	Logistica	RF	predi133	Logistica	RF	SVMLinear
predi109	Logistica	GB(gbm)	predi134	Logistica	GB(gbm)	GB(Xgboost)
predi110	Logistica	GB(Xgboost)	predi136	Logistica	GB(gbm)	SVMPoly
predi111	Logistica	SVMPoli	predi137	Logistica	GB(gbm)	SVMLinear
predi112	Logistica	SVMLinear	predi138	Logistica	GB(Xgboost)	SVMPoly
predi113	Red	RF	predi139	Logistica	GB(Xgboost)	SVMLinear
predi114	Red	GB(gbm)	predi140	RF	GB(gbm)	SVMPoly
predi115	Red	GB(Xgboost)	predi141	RF	GB(gbm)	SVMLinear
predi116	Red	SVMPoli	predi142	RF	GB(Xgboost)	SVMPoly
predi117	Red	SVMLinear	predi143	RF	GB(Xgboost)	SVMLinear
predi118	RF	GB(gbm)	predi144	RF	Red	GB(gbm)
predi119	RF	GB(Xgboost)	predi145	RF	Red	GB(Xgboost)
predi120	RF	SVMPoli	predi146	RF	Red	SVMPoly
predi121	RF	SVMLinear	predi147	RF	Red	SVMLinear
predi122	GB(gbm)	GB(Xgboost)	predi148	Red	GB(gbm)	SVMPoly
predi123	GB(gbm)	SVMPoli	predi149	Red	GB(gbm)	SVMLinear
predi124	GB(gbm)	SVMLinear	predi150	Logistica	RF	GB(gbm)
			predi151	Logistica	RF	GB(Xgboost)
			predi153	Logistica	RF	GB(Xgboost)
			predi154	Logistica	RF	GB(Xgboost)

El resultado y las conclusiones del ensamblado en R se parecen mucho a los de SAS: las técnicas de ensamblado dan lugar a mejoras en la tasa de error medio, pero cabe valorar si estas mejoras compensan la complejidad que aporta este proceso a los modelos. El mejor modelo de ensamblado es el ***predi109*** obtenido a partir de la combinación de los dos mejores modelos previamente comentados: la logística y *gradient boosting*. Además, observamos que una buena alternativa también es ***predi204***, combinación de *SVM* lineal y *gradient boosting* que consigue reducir la varianza del primer modelo de ensamblado. A pesar de ello, si nos fijamos en el modelo simple de *gradient boosting gbm (nombre gbm en el gráfico)* vemos como no se producen mejoras importantes en sesgo ni varianza, por lo que concluimos que el ensamblado no aporta mejoras importantes y no compensa llevarlo a cabo en este estudio.

5.10 Exploración de la no tramificación de las variables

Finalmente para completar el análisis se ha llevado a cabo una exploración alternativa de modelización siguiendo el mismo proceso de preparación de los datos, depuración, selección de variables y prueba de algoritmos pero sin llevar a cabo el paso de tramificación WOE de las variables, es decir, manteniendo sus forma, estructura y valor original. El desarrollo de la modelización es prácticamente la misma, solo que al llegar al punto de tramificación y asociación de valor WOE a cada uno de los tramos generados simplemente no se hace y se mantienen las variables en su estado original. De esta forma sí tendremos variables categóricas y continuas en el estudio, para las cuales tendremos que seguir otros criterios de selección y tratamiento para la posterior modelización.

Para este proceso hemos realizado un nuevo filtrado de las variables, por lo que el set final utilizado no coincide completamente con el que obtenemos para la modelización con tramificación. A continuación en la figura 39 se incluye una tabla comparativa con las variables usadas en cada caso:

Figura 39: Tabla Resumen Sets para Modelización tramificando vs no tramificando

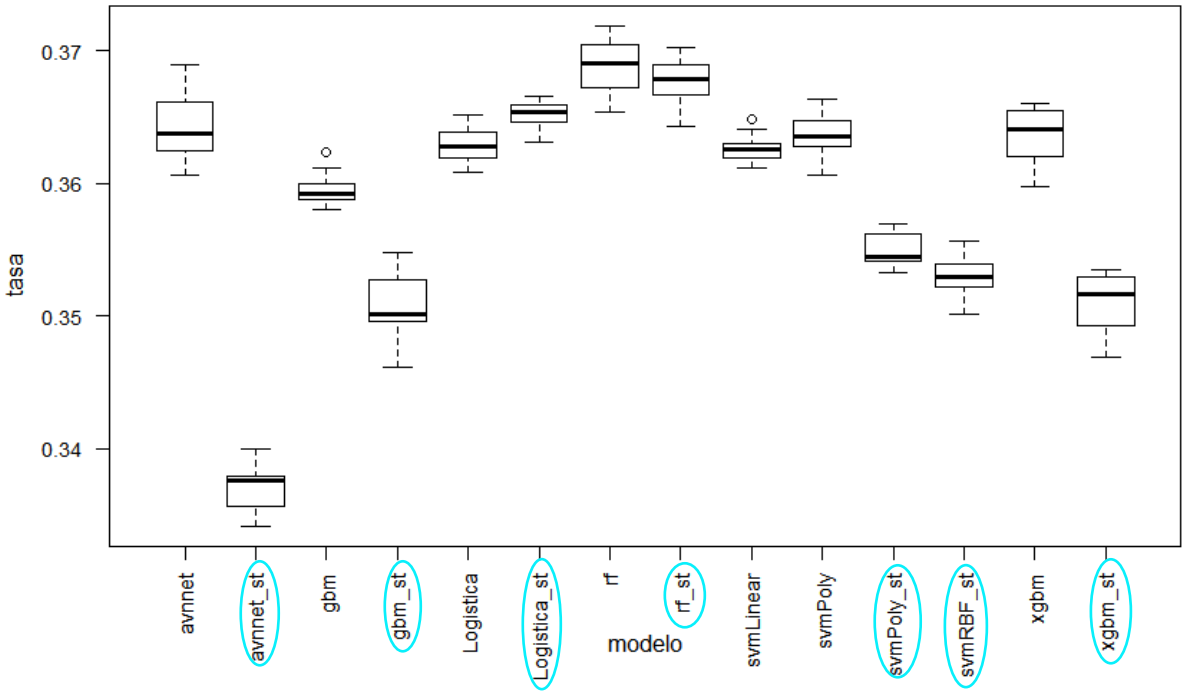
Variables Tramificado		Variables sin tramificar	
		Variable	**Tipo**
all_util	X5	*no usada*	-
both_inq_last_6mths	X6	both_inq_last_6mths	intervalo
dti_total	X8	*no usada*	-
installment	X11	installment	intervalo
tot_hi_cred_lim	X16	tot_hi_cred_lim	intervalo
total_bc_limit	X17	total_bc_limit	intervalo
addr_state	X19	D_addr_state1	dummy
		D_addr_state4	dummy
emp_length	X21	D_emp_length1	dummy
home_ownership	X22	D_home_ownership2	dummy
mths_since_recent_inq	X23	D_mths_since_recent_inq4	dummy
purpose	X24	D_purpose4	dummy
term	X25	D_term1	dummy
verification_status_joint	X26	D_verification_status_join1	dummy
SQRT_loan_amnt	X3	SQRT_loan_amnt	intervalo
nom_tot_cur_bal	X14	*no usada*	-
acc_open_past_24mths	X4	acc_open_past_24mths	intervalo
il_util	X9	*no usada*	-
total_rev_hi_lim	X18	total_rev_hi_lim	intervalo
no usada	-	inq_fi	intervalo
no usada	-	nom_revol_bal	intervalo

Es importante comprobar que se trata de sets similares y que tiene sentido comparar los resultados para cada algoritmo dependiendo de si se usa un set de datos u otro para evitar conclusiones erróneas. Efectivamente, vemos que son equiparables ya que la estructura es muy parecida y en lugar de incluir 18 variables como hemos utilizado con la tramificación, ahora se incluyen 17 y además hay otras variaciones en la composición. Hay un total de 14 variables que coinciden y observamos que principalmente son las que en el ranking presentaban consenso de uso en todos los modelos, por lo que el siguiente paso ha sido verificar si para las que no coinciden existen otras en el set con un aporte de información equivalente. Tras confirmar que así es, concluimos que es oportuno proceder con la evaluación y comparación de los modelos generados para cada algoritmo según se use el set con o sin tramificación.

Otro aspecto importante que hemos tenido en cuenta con respecto a las diferencias entre el set de variables tramificadas es que dejamos de modelizar solamente con variables continuas y tenemos también variables categóricas. Para los modelos hemos tenido que tratar de forma diferente a estas variables mediante la realización dummies para cada una de sus categorías y utilizar el filtro de importancia que permite mantener únicamente aquellos segmentos relevantes. De esta forma evitamos sobre parametrizar el modelo y dejamos exclusivamente la información que procede mantener. En la tabla anterior se indica cuáles son estas variables y cuántas categorías dummy mantenemos para cada una.

Finalmente procedemos con la comparación de cada uno de los algoritmos según se utilice un set de datos u otro, evaluados el nivel de tasa de fallo mediante validación cruzada repetida en R. Para ello utilizamos el siguiente diagrama de cajas, el cual muestra los mejores modelos obtenidos para cada set. Para los datos tramificados usamos la configuración previamente presentada, mientras que para el set de datos no tramificado incluimos la configuración en el anexo V:

Figura 40: Diagrama de Cajas Comparación de modelos con y sin tramificación



De este análisis podemos extraer las siguientes conclusiones:

1. El mejor modelo para el set sin tramificar es la red neuronal, mientras que para el set de datos no tramificado ya hemos visto que no era precisamente de las mejores opciones (resultado comparado en R). Por ello extraemos que las redes neuronales captan mejor las relaciones de los datos con la variable objetivo sin tramificar.
2. El modelo *gradient boosting* sin tramificar es mejor que tramificado en términos de tasa de error medio. Esto es así para ambos paquetes que permiten configurar este modelo en R: *gbm* y *xgboost*. A diferencia de lo que hemos visto anteriormente al evaluar el mejor modelo con los datos tramificados, *gradient boosting* era el mejor modelo en términos de error, mientras que ahora es superado por la red neuronal.
3. La regresión logística funciona algo mejor con las variables tramificadas pero realmente la diferencia es muy pequeña y podría venir generada por las variables de menos que se incluyen en el set de datos sin tramificar.
4. El modelo *random forest*, como ya hemos detallado antes, de todos los que hemos probado es el menos apropiado para modelizar estos datos y esta conclusión aplica tanto para el set con variables tramificadas como sin tramificar.
5. En cuanto al algoritmo *SVM*, vemos como se adaptan mejor al set de datos sin tramificar, consiguiendo un error menor. Además, para el set de datos sin tramificar los mejores *kernel* son el polinomial y el radial (*RBF*), captando un tipo de relación no lineal que no se consigue captar con el set de datos tramificado (para el cual ya hemos visto previamente que en R la mejor alternativa es el *kernel* lineal y un polinomial sencillo con un grado dos).
6. En general, la tasa de error obtenida sin tramificar es mejor que utilizando tramificación. La diferencia más grande se aprecia entre el mejor algoritmo obtenido para el set de datos tramificado (*gbm*) y el mejor algoritmo obtenido para el set de datos sin tramificar (*avnnet_st*). Se trata de una tasa de error un 2% más baja, lo cual podría hacernos elegir el set como mejor opción . Por otra parte, también pensamos que basarnos solamente en este criterio para decidir cuál es la mejor alternativa podría llevarnos a conclusiones erróneas debido a que para el resto de los casos, la mejora media del error producida según se utilice un set u otro está en torno al 1%. Por ello es muy importante ver según el algoritmo que nos interese utilizar para modelizar qué set de datos puede permitir obtener mejores resultados. Por ejemplo, si nos centramos en la red neuronal en R, claramente es el mejor algoritmo para modelizar los datos sin tramificar (*avnnet_st*) y se aprecia una diferencia relevante con la red que generamos para el set de datos con tramificación (*avnnet*). En este caso la mejor alternativa es usar el set de datos no tramificado.
7. Cabe mencionar que al utilizar el set de datos tramificado hemos experimentado mayor velocidad de ejecución, por lo que es un factor que puede influir en la decisión de tramificar o no las variables. Además, si finalmente decidimos que la regresión logística es la mejor opción por el hecho de que el algoritmo clásico presenta la ventaja de mantenimiento de interpretación de los resultados con respecto a las cajas negras de los algoritmos de *Machine Learning*, podríamos decidir utilizar el set de datos tramificado, ya que la división en categorías es más clara y permite una interpretación más visual.

Dados todos estos argumentos, no podemos decidir que sea siempre mejor tramificar o no tramificar los datos, si no que depende del criterio en el que nos basemos y de los

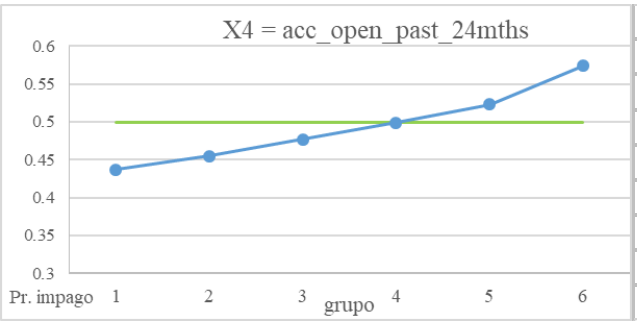
algoritmos que nos interese utilizar escogeremos una u otra alternativa. Para el set de datos sin tramificar cabría valorar si es mejor modelizar con Gradient Boosting o con la Logística, evaluando si la mejora producida en el error compensa la pérdida de interpretabilidad de los resultados que permite la logística.

5.11 Efectos Marginales de la Regresión

Finalmente, tras haber realizado pruebas con diversos algoritmos, modelos clásicos, utilizando diferentes técnicas, configuraciones de parámetros e incluso utilizando dos programas distintos, podemos concluir que la regresión logística termina siendo el algoritmo más completo por mantener una relación sesgo-varianza competitiva respecto a los algoritmos de *Machine Learning* y además aporta interpretación para entender los resultados que estamos obteniendo. Dicho esto, vamos a utilizar la mejor regresión logística conseguida previamente, la cual seleccionaba mediante un proceso iterativo un total de 15 variables, para realizar un estudio de los efectos marginales para cada una de ellas.

Este proceso consiste en calcular un individuo medio ficticio, es decir, una observación que toma el valor medio en todas las variables que incluye la regresión. A continuación construimos una matriz de datos en la que mantenemos para todas las variables del estudio el valor medio y solamente mantenemos el valor original del tramo de la variable que queramos evaluar. Tenemos que generar tantas filas como categorías de variables tengamos por analizar. Finalmente, lo que hacemos es utilizar la mejor regresión que hemos obtenido para realizar predicciones sobre cómo va variando la probabilidad de impago para cada uno de estos individuos dependiendo de si, para cada variable, se caracterizan por estar en un tramo u otro. A continuación representamos gráficamente este efecto para el set de modelización adaptado a la regresión, aportando interpretabilidad y una medida de puntuación alternativa de referencia:

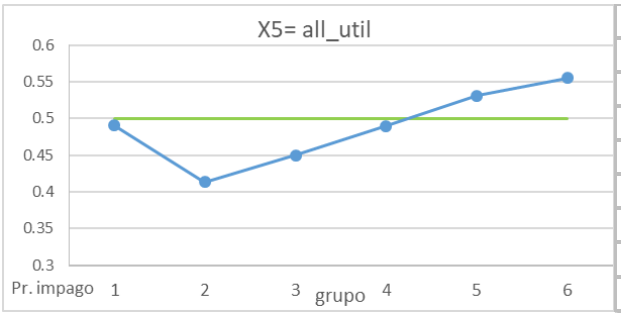
Figura 41: Efecto Marginal acc_open_past_24mths



X4			
Grupo		Predicción	Incremento
1	0 -1	0.4367439	
2	2	0.4548979	1.82%
3	3	0.4768377	2.19%
4	4,6	0.4990621	2.22%
5	6,7	0.523026	2.40%
6	8<	0.5739224	5.09%
Individuo Medio = 0.5			

Para la variable número de cuentas de crédito abiertas en los últimos 24 meses vemos cómo a medida que aumenta la cantidad, aumenta el porcentaje de impago de forma progresiva. La línea verde representa el individuo medio y el punto de corte, es decir, a partir de qué umbral determinamos con esta variable que este individuo toma características que tiendan a generar impago. Así observamos como el incremento se vuelve mayor a medida que nos situamos en torno a mayor cantidad de cuentas. Puede ser de gran utilidad para la empresa considerar la cantidad de líneas de crédito recientemente abiertas para evaluar el riesgo del individuo.

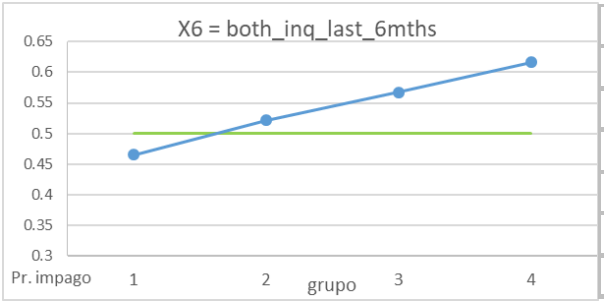
Figura 42: Efecto marginal all_util



X5			
Grupo		Predicción	Incremento
1	0-31	0.4908236	
2	31-37	0.4136854	-7.71%
3	37-44	0.4495911	3.59%
4	44-65	0.4898864	4.03%
5	65-82	0.5312785	4.14%
6	82<	0.5552999	2.40%
Individuo Medio = 0.5			

Esta variable es un ratio calculado con balance total del límite de crédito partido el balance de todas las cuentas. Como podemos observar, tener un ratio muy pequeño entre 0-31 presenta más riesgo que tener un ratio entre 31-37, lo cual es normal debido a que aquellos individuos con poca disponibilidad de cantidad para endeudarse normalmente es porque se identifica cierto riesgo en ellos y el margen de endeudamiento cedido es menor. Dicho esto, se trata de individuos con un ratio menor, es decir, situados en la categoría 1 y además con un porcentaje de impago algo más elevado que los individuos correspondientes a la categoría 2, los cuales siguen presentando ratios pequeños pero no tanto por tener un límite de crédito pequeño, sino por tener buen ratio de endeudamiento. Claramente vemos como el riesgo de impago comienza a dispararse cuando se supera el 50% del ratio (categorías 4/5) y finalmente el incremento se vuelve más gradual al superar el 62% de endeudamiento.

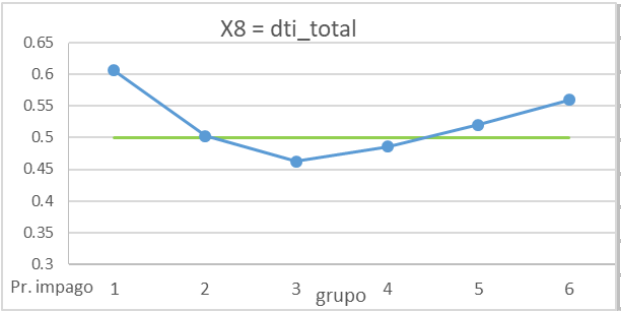
Figura 43: Efecto Marginal both_inq_last_6mths



X6			
Grupo		Predicción	Incremento
1	0,1	0.4651156	
2	1,2	0.5217819	5.67%
3	2,3	0.5669395	4.52%
4	3<	0.6165632	4.96%
Individuo Medio = 0.5			

Esta variable refleja el número de consultas que se han tenido que realizar para el individuo en el registro público en los últimos 6 meses. Como podemos observar, a más consultas más probabilidad de impago. Entre 0 y 1 consulta (grupo 1) el riesgo es menor pero a partir de 2 ya se incrementa en casi un 6% la probabilidad de impago. El porcentaje de incremento se mantiene cercano al 5% para cantidades de más de 3 consultas.

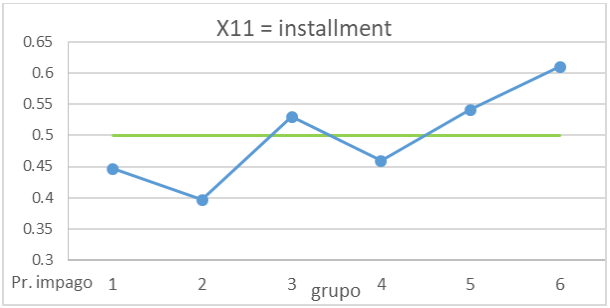
Figura 44: Efecto Marginal dti_total



X8			
Grupo		Predicción	Incremento
1	<3.42	0.6064783	
2	3.42-7.79	0.5029052	-10.36%
3	7.79-13.59	0.4625744	-4.03%
4	13.59-24.29	0.4859154	2.33%
5	24.29-33.96	0.5203141	3.44%
6	33.96<	0.5599303	3.96%
Individuo Medio = 0.5			

Esta variable es un ratio calculado como pago mensual de deuda por obligaciones partido renta mensual reportada. Al tratarse de ratios de menos de 3.42% la probabilidad de impago es muy elevada debido a que el pago mensual de deudas es por cantidades pequeñas, lo cual suele ir asociado a individuos que no pueden hacer frente a cantidades superiores dado su nivel de renta. Para niveles intermedios la probabilidad disminuye y es a partir de ratios superiores al 24% donde observamos como vuelve a subir la probabilidad de impago de nuevo, en este caso por presentar pagos mensuales de considerable tamaño en proporción a la renta mensual reportada, lo cual también es un claro indicador de riesgo de impago. Ratios muy pequeños presentan mayor riesgo que ratios grandes en este caso.

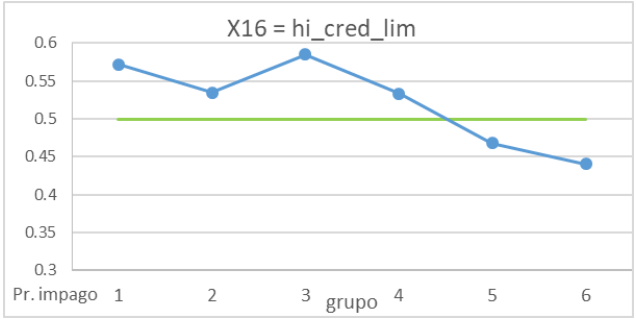
Figura 45: Efecto Marginal installment



X11			
Grupo		Predicción	Incremento
1	<298.29	0.4466552	
2	298.29-341.93	0.3964462	-5.02%
3	341.93-402.98	0.5302687	13.38%
4	402.98-495.61	0.4591688	-7.11%
5	495.61-830.12	0.5414915	8.23%
6	830.12<	0.6103528	6.89%
Individuo Medio = 0.5			

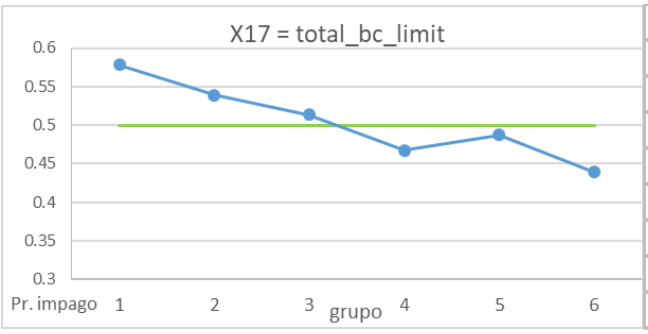
Esta variable refleja la cantidad mensual a pagar por el crédito. Está calculada teniendo en cuenta la cantidad del crédito, tipo de interés y plazos. Por ello, vemos como para los créditos con menor cantidad a pagar, la probabilidad de impago es menor que para aquellos créditos con mayor cantidad mensual a pagar. Tiene sentido que la relación no sea lineal ya que, en muchos casos, cantidades más pequeñas (grupo 1) pueden llevar un mayor riesgo asociado por tratarse de solicitantes con créditos de menor cantidad por tener mayor riesgo por sus características que por ello se les conceda préstamos de menor importe. En el grupo 2 claramente encontramos los créditos con menor riesgo de impago, mientras que en el grupo 3 este se dispara. En el grupo 4 encontramos importes algo más elevados pero la probabilidad de impago de estos disminuye de nuevo, mientras que ya para cantidades superiores observamos como el riesgo aumenta. De aquí extraemos que esta variable contiene una relación no lineal pero con un aporte e interpretación interesante asociada a la información implícita de todos los componentes que permiten crear esta variable. Queda claro que, a mayor cantidad de obligación mensual, mayor riesgo de impago asociado.

Figura 46: Efecto Marginal hi_cred_lim



X16			
Grupo		Predicción	Incremento
1	<31700	0.5722048	
2	31700-46006	0.5344442	-3.78%
3	46006-53975	0.5847544	5.03%
4	53975-99144	0.5332768	-5.15%
5	99144-174822	0.468215	-6.51%
6	174822<	0.4406875	-2.75%
Individuo Medio = 0.5			

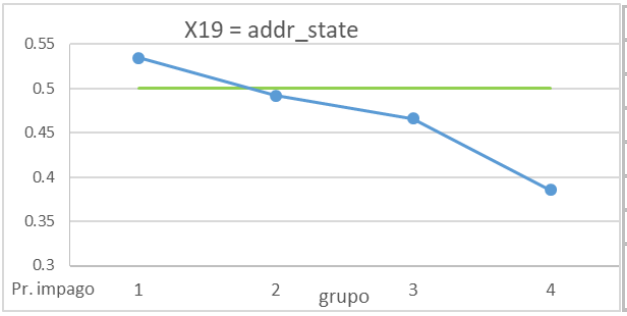
Figura 47: Efecto Marginal total_bc_limit



X17			
Grupo		Predicción	Incremento
1	<5000	0.5781842	
2	5000-6500	0.5389244	-3.93%
3	6500-18100	0.5134607	-2.55%
4	18100-26000	0.4674243	-4.60%
5	26000-34300	0.4875419	2.01%
6	>34300	0.4391425	-4.84%
Individuo Medio = 0.5			

Estas dos variables muestran el máximo que se pueden endeudar en todas las cuentas en general y para las tarjetas de crédito. Como podemos observar, la relación es inversa para ambas y a menos cantidad disponible, más elevado es el riesgo. Esto es debido a que un individuo al que se le da mayor margen de endeudamiento es por presentar características que indican capacidad de afrontar sus deudas y por lo tanto menor peligro de impago. Por otra parte, los individuos con mayor restricción de crédito suelen asociarse a características que implican mayor probabilidad de impago, de aquí esta relación decreciente.

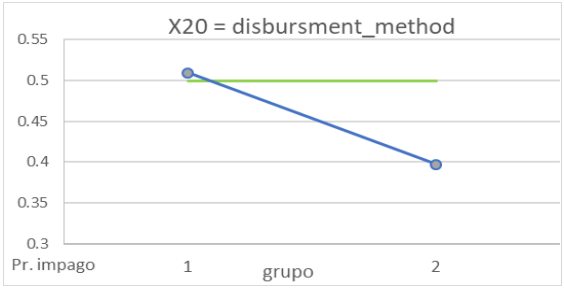
Figura 48: Efecto marginal addr_state



X19			
Grupo		Predicción	Incremento
1	4*	0.5342301	
2	3*	0.4919791	-4.23%
3	2*	0.4661686	-2.58%
4	1*	0.3856803	-8.05%
Individuo Medio = 0.5			
*La composición de cada grupo aparece al principio del trabajo			

Esta variable es la agrupación de los 50 Estados de EEUA realizada previamente con un árbol para realizar grupos con niveles de impago parecidos. La composición de cada uno de ellos se incluye al principio del trabajo y podría ser útil para asociar mayor probabilidad de impago a las diferentes zonas del país.

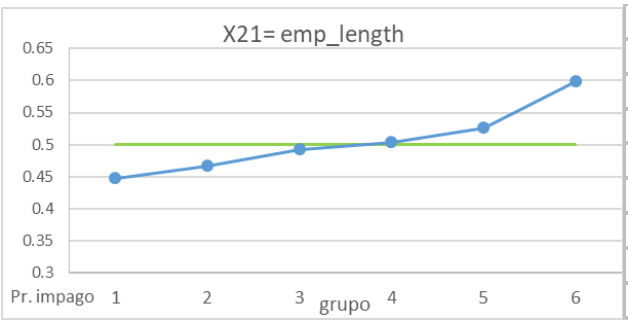
Figura 49: Efecto Marginal disbursment_method



X20			
Grupo		Predicción	Incremento
1	Cash	0.39719	
2	Direct Pay	0.5090142	11.18%
Individuo Medio = 0.5			

Esta variable representa como según si el dinero del crédito se adquiere mediante un pago directo o en efectivo puede afectar a la probabilidad de impago del individuo. Los créditos recibidos mediante pago directo tienen una probabilidad de impago el 11% superior.

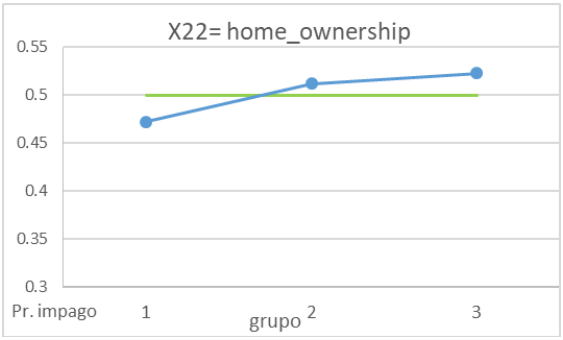
Figura 50: Efecto Marginal emp_length



X21			
Grupo		Predicción	Incremento
1	6 ,9 años	0.4479257	
2	>10 años	0.4668696	1.89%
3	2,5,7 años	0.4930236	2.62%
4	1,4 años	0.5035434	1.05%
5	3,8, <1 años	0.5265744	2.30%
6	N/A	0.5986669	7.21%
Individuo Medio = 0.5			

Esta variable representa el tiempo que llevan los individuos trabajando. Como podemos observar, el menor riesgo se asocia a tener más antigüedad y para aquellos individuos que tengan antigüedad menor a 4 años o que se caractericen como *na* (no aplica) presentan los porcentajes más elevados. Especialmente este último grupo lleva asociado un riesgo de impago muy elevado en relación a la media; por ello deducimos que este grupo contiene aquellos individuos que quizá no tienen trabajo pero por tener otras rentas o respaldo para solicitar un crédito han podido adquirirlo. Además, extraemos que a más estabilidad laboral (más años de contrato), menor es el riesgo de impago.

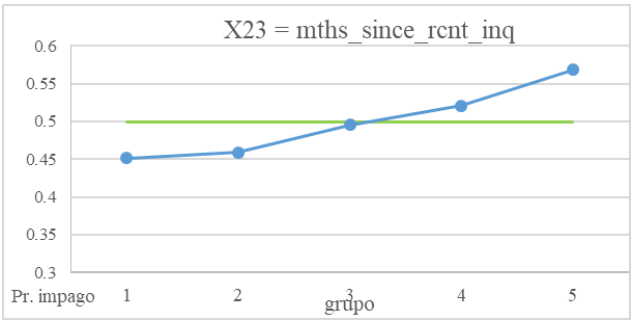
Figura 51: Efecto Marginal home_ownership



X22			
Grupo		Predicción	Incremento
1	Hipoteca/otros	0.4725848	
2	En propiedad	0.5118732	3.93%
3	Alquiler	0.5230429	1.12%
Individuo Medio = 0.5			

En este gráfico representamos como varía la probabilidad de impago según sea el tipo de relación contractual en cuanto a la residencia habitual. Aquellos individuos que han adquirido su propia casa mediante hipoteca o su estado de residencia es otro no especificado, presentan menor probabilidad de impago. Por otra parte, aquellos que están de alquiler tienen mayor probabilidad de impago de sus obligaciones.

Figura 52: Efecto Marginal mths_since_rcnt_inq

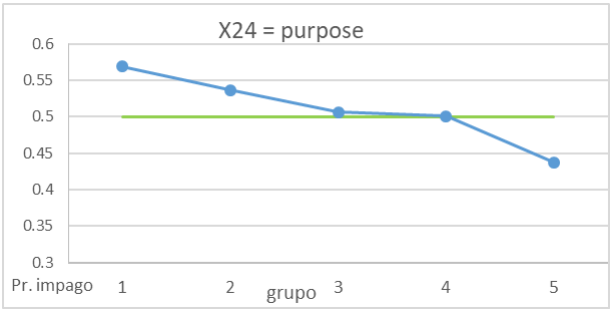


X23			
Grupo		Predicción	Incremento
1	4	0.4510837	
2	3	0.4593214	0.82%
3	2	0.4956415	3.63%
4	1	0.5208697	2.52%
5	0	0.5682514	4.74%
Individuo Medio = 0.5			

Esta variable representa el número de meses transcurridos desde la última consulta pública que se ha realizado a la persona. Como podemos observar, cuantos más meses hace que han consultado a la persona (grupo 1), menos probabilidad de impago. Esta va creciendo de forma gradual a medida que se vuelve más reciente la última fecha de

consulta, siendo la probabilidad más alta para aquellos individuos a los que hace cero meses que se les ha tenido que consultar (grupo 5).

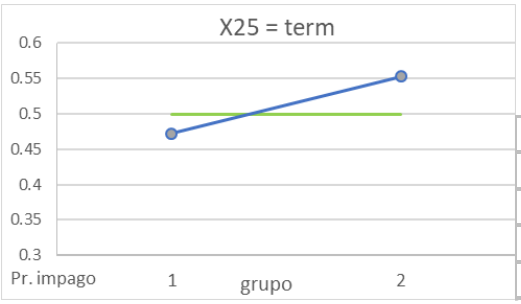
Figura 53: Efecto Marginal purpose



X24			
Grupo		Predicción	Incremento
1	major_purchase, medical, small_business	0.5689857	
2	house, moving, others	0.5364941	-3.25%
3	debt_consolidation	0.5063745	50.64%
4	car, home_improvement, vacation	0.5008858	-3.56%
5	credit_card, renewable_energy	0.4374649	-6.89%
Individuo Medio = 0.5			

Esta variable muestra según el uso que se le da al crédito, el riesgo asociado. Para aquellos préstamos destinados a compras por importe elevado, uso médico y de negocios pequeños son a los que se asocia una mayor probabilidad de impago, seguidos por aquellas destinadas al hogar, mudanzas y otros usos varios. Los usos más seguros son los créditos utilizados para disminuir deuda de tarjetas de crédito y aquellos destinados a proyectos de energías renovables.

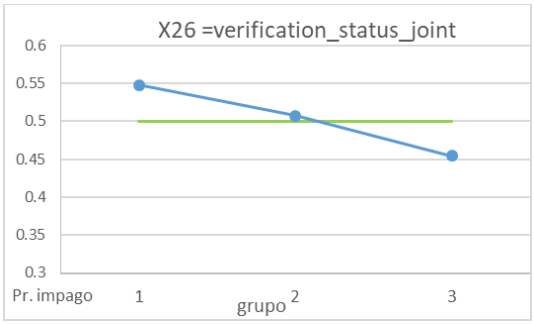
Figura 54: Efecto Marginal term



X25			
Grupo		Predicción	Incremento
2	36 months	0.4719582	
1	60 months	0.5531883	8.12%
Individuo Medio = 0.5			

Esta variable muestra el plazo de devolución del préstamo establecido; este puede ser a 36 o a 60 meses. Cuanto mayor es el plazo, más riesgo de impago asociado y la probabilidad de impago aumenta en un 8%.

Figura 55: Efecto Marginal verification_status_joint



X26			
Grupo		Predicción	Incremento
1	Verificado	0.5480207	
2	Fuente no verificada	0.5071229	-4.09%
3	No verificado	0.454392	-5.27%
Individuo Medio = 0.5			

Esta variable representa el estado de verificación de los ingresos de los solicitantes. Si se verificado, la probabilidad de impago es más alta frente a los casos en los que ha sido necesario, lo cual lleva a pensar que si se ha tenido que verificar ha sido debido a otras características que han hecho dudar de la capacidad de afrontar la deuda del individuo.

Finalmente, tras realizar este análisis tenemos una idea orientativa de qué cualidades pueden ayudar a la empresa a determinar si es apropiado o no conceder un crédito a un

individuo. Los modelos de *scoring* se basan en asociar puntos a cada uno de estos tramos y realizar un cómputo del total para cada uno de los solicitantes según se comporten en cada variable. Así se consigue detectar individuos con ciertas características personales ajenas a la empresa asociadas a elevada probabilidad de presentar dificultades en los pagos futuros del crédito solicitado. Además, es importante establecer umbrales de riesgo adecuados dispuestos a asumir por la empresa y que ellos determinen puntos de corte a partir de cierta puntuación a partir de la cual se decida no conceder los préstamos. Esta decisión es muy importante, ya que de ella depende que los inversores puedan generar rentabilidades elevadas a los que desean especular, pero a su vez siendo realistas y evitando que se invierta en vano.

6. Conclusiones del trabajo y posibles líneas futuras de investigación.

El objetivo del trabajo era, por un lado, encontrar el mejor algoritmo que nos permitiera elaborar un modelo de detección de impago alternativo al que actualmente utiliza LendingClub para asegurar a los inversores su recobro y rentabilidad, permitiendo así mantener el negocio en funcionamiento. Es importante contar con una herramienta precisa para detectar las solicitudes que con elevada probabilidad incumplirán las condiciones y generarán pérdidas para los inversores y la empresa para evitar concederles un préstamo.

Para alcanzar la mejor alternativa de modelización hemos tenido que llevar a cabo un estudio con muchos pasos, de los cuales hemos ido extrayendo diferentes puntos clave a tener en cuenta para sustentar el trabajo. Primero de todo, vemos como de todas las variables que facilitaba la empresa se debe llevar a cabo un procedimiento de preparación y composición del set de datos para poder comenzar a utilizarlos y aplicarlos a la modelización. Dados los estudios de las variables de forma individual y de su aporte conjuntamente al set, determinamos que sería conveniente agilizar el sistema de solicitud y aprobación del préstamo. Esto puede conseguirse mediante una simplificación del cuestionario y requiriendo solamente aquella información realmente útil, preguntando la información de una forma algo más concreta y alineada con el objetivo del trabajo. Además, se podría estandarizar cierta información en la solicitud como el área de profesión para que pudiera generar aporte útil a la modelización. Se podría valorar la opción de considerar si otras variables algo más personales y relacionadas con el individuo podrían aportar valor al modelo (edad, estudios, situación familiar etc.). También debería considerarse es realizar dos tipos de bases de datos según sean las solicitudes individuales o conjuntas para realizar modelos diferentes según para cada una de ellas y evitar problemas con las variables que aplican solamente a solicitudes conjuntas y viceversa.

Ya adentrándonos en la modelización, vemos como reduciendo la cantidad de variables conseguimos facilitar un criterio de puntuación alternativo al rating que ofrece LendingClub para que los inversores compongan su cartera. Nosotros finalmente pensamos que la mejor alternativa es utilizar la regresión logística con 15 variables input tramificadas, lo cual permite realizar una tarjeta de puntuación y, a partir del *scoring* asociado a cada tramo, conseguir determinar la probabilidad total impago asociada a cada individuo. Además, para evitar quedarnos con la duda de si al tramificar los datos estamos favoreciendo la logística hemos realizado un estudio paralelo sin tramificar las variables con el que llegamos a la misma conclusión: una vez más las técnicas más complejas de

Machine Learning no consiguen resultados en la tasa de fallo lo suficientemente mejores que los de la regresión logística. La mejora que deberíamos observar en el error debería compensar la pérdida de interpretabilidad de las variables del modelo, ya que la regresión permite indagar en él, ver el papel que tiene cada variable y alinear las diferencias en la caracterización de los individuos a su categorización como futuros buenos o malos pagadores. Al haber realizado pruebas en diferentes softwares y haber llegado a resultados similares podemos finalmente confirmar con mayor seguridad que la mejor alternativa es utilizar la regresión logística para desarrollar el modelo de *credit scoring* para LendingClub.

Otro aspecto que cabe comentar es que las variables que acabamos recogiendo en el modelo aportan el mismo tipo de información que en la introducción del trabajo hemos especificado que LendingClub indica tener en cuenta para valorar las solicitudes: historial de pagos, antigüedad y tipo de créditos que solicita, porcentaje de uso dado el límite total, balance total o balance de deuda, comportamiento reciente de crédito, consultas realizadas al individuo y disponibilidad de crédito.

Finalmente, como posibles recomendaciones que podríamos dar a la empresa serían que centraran sus intereses en solicitar información más sencilla y estandarizada, permitiendo así a los inversores interesados llevar a cabo sus propios criterios de valoración y desarrollar sus propios estudios alternativos sin tener que conformarse con la puntuación ofrecida por la propia empresa. Además, podría ser interesante que ofrecieran más información sobre la metodología utilizada y sobre los criterios en los que se basan para determinar si se concede o no un préstamo y, en caso de que se conceda, facilitar también más información sobre el modelo de determinación de riesgo. De esta forma los inversores también podrían valorar si quieren basarse exactamente en los mismos criterios o si prefieren aplicar alguna que otra variante para determinar la composición de sus carteras. Una posible evolución de este trabajo podría ser a partir de la regresión que hemos obtenido crear criterios de corte para las probabilidades de impago y generar un *rating* que pueda compararse con la puntuación facilitada por la empresa a cada solicitud. Podría ser una herramienta útil alternativa para aquellos inversores que busquen depositar sus ahorros basándose en un criterio propio de puntuación y poder evaluar si la relación rentabilidad riesgo establecida por la empresa es la adecuada a cada porción que se están planteando adquirir para componer su cartera de inversión.

7. Bibliografía

BRUNO, A. (18 de marzo de 2017). *The 2008 Financial Crisis Explained.* Obtenido de https://www.lombardiletter.com/the-2008-financial-crisis-explained/11672/

COMPANY BLASCO, V. 2018. *Modelo de Clasificación para Inversión en Préstamos de Bondora.* L. Escot Mangas (dir.) Trabajo Final de Master, Universidad Complutense de Madrid

DE LA CRUZ, M. (17 de Septiembre de 2018). Nueve de cada diez partícipes de fondos conservadores, en pérdidas este año. *Expansión*. Disponible en: http://www.expansion.com/ahorro/2018/09/17/5b9b9466ca4741f55d8b4591.html

HAOTIAN, C., ZIYUAN, C., TIANYU, X. y ZHOU, Y., 2015. *Data Mining on Loan Default Prediction*. Boston College.

LendingClub centro de ayuda. https://www.lendingclub.com/

PORTELA, J. 2019. *Apuntes Machine Learning*. Facultad de Estudios Estadísticos, Universidad Complutensde de Madrid

SAS Technical Support, s.f. *Overview: LOGISTIC Procedure.* SAS. Disponible en: https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect001.htm

SADDIQI, N., 2005. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* SAS Institute Inc.

SKANTZOA, N. y CATELEIN, N., 2016. *Credit Scoring Case study in data analytics*.

VALLE CARRASCAL, J. M. 2015. *Modelos de medición del riesgo de crédito.* A. Fernández Ruiz y M. Elices López (dir.). Tesis Doctoral, Universidad Complutense de Madrid.

8. Anexos

Anexo I - Descripción del set de datos original

Variable	Definición	Categoría	Subcategoría	Tipo de Variable
id	Número identificador creado para cada registro	solicitud	identificadora	id
member_id	Número identificador de LC para el individuo	solicitante	identificadora	id
loan_amnt	Cantidad de crédito solicitada	solicitud	importe	intervalo
funded_amnt	Cantidad del crédito solicitado concedida en la actualidad	solicitud	importe	intervalo
funded_amnt_inv	Cantidad de crédito en la que se ha invertido en el momento actual	préstamo	balance	intervalo
term	Número de pagos (en meses) elegidos para devolver el crédito: toma valores 36 o 60 meses	préstamo	caracterizadora	categórica (binaria)
int_rate	Tipo de interés asociado al crédito	préstamo	caracterizadora	intervalo
installment	Cantidad mensual a pagar por el crédito	préstamo	pagos	intervalo
grade	Calificación asignada por LC	préstamo	calificación	categorica
sub_grade	Sub-calificación asignada por LC	préstamo	calificación	categorica
emp_title	Descripción del tipo de empleo desempeñado por el prestatario principal	solicitante	social	categorica
emp_length	Tiempo que lleva trabajando el prestatario principal en años. Pueden ser: <1 year, 1 year, 2 years ... 9 years, 10 + years, n/a	solicitante	social	categorica
home_ownership	Caracterización del status del individuo respecto a su vivienda actual: RENT, OWN, MORTGAGE, ANY	solicitante	social	categorica
annual_inc	Renta Individual declarada por el individuo principal en el momento de solicitud	solicitante	económica	intervalo
verification_status	Indica si los ingresos han sido verificados, no verificados o si se ha verificado la fuente de ingresos	solicitante	económica	categorica
issue_d	Mes en el que se concedió el crédito: abril, mayo, junio	préstamo	caracterizadora	categorica
loan_status	Estado actual del préstamo. Puede ser: current, charged off, default, fully paid, in grace period, late (16-30 days) late (31-120 days)	préstamo	estado	categorica
pymnt_plan	Indica si se ha requerido a fecha actual un plan de reestructuración de pagos	préstamo	reestructuración	categorica (binaria)
url	URL de la página de LC con listado de información	préstamo	caracterizadora	id
desc	Detalles facilitados sobre el uso del crédito (opcional)	préstamo	caracterizadora	texto
purpose	Uso que se le dará al crédito: car, credit_card, debt_consolidation, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation	préstamo	caracterizadora	categorica
title	Descripción del uso del crédito facilitada por el prestatario: car, credit_card, debt_consolidation, home_improvement, house, major_purchase, medical, moving, other, renewable_energy, small_business, vacation	préstamo	caracterizadora	categorica
zip_code	Tres primeros dígitos del código postal	solicitante	social	categorica
addr_state	Estado del solicitante (50 estados de Estados Unidos de America)	solicitante	social	categorica
dti	Ratio calculado a partir del pago mensual de deuda sobre el total de deuda,	préstamo	balance	intervalo
delinq_2yrs	Número de vencimientos de + de 30 días (delincuencias) que aparecen en el registro dentro de los últimos 2 años	solicitante	impago	categorica
earliest_cr_line	Mes en el que está reportado que se abrió la primera línea de crédito	solicitante	cuentas	categorica
inq_last_6mths	Número de consultas que han tenido que realizarse al registro público de crédito durante los últimos 6 meses	solicitante	consultas	categorica
mths_since_last_delinq	Número de meses transcurridos desde la última delincuencia del individuo	solicitante	impago	categorica
mths_since_last_record	Número de meses transcurridos desde el último registro público	solicitante	consultas	categorica
open_acc	Número de lineas de credito abiertas que aparecen en el reporte de crédito	solicitante	cuentas	categorica
pub_rec	Número de cuentas derogatorias que aparecen en el registro de crédito	solicitante	impago	categorica
revol_bal	Balance total de cuentas de tipo "revolving"	solicitante	balance	intervalo
revol_util	Ratio: utilización de la línea de crédito de tipo "revolving" / cantidad de crédito de tipo "revolving" que el individuo está utilizando dado el total disponible para utilizar	solicitante	balance	intervalo
total_acc	Total líneas de crédito contenidas en el archivo de crédito	solicitante	cuentas	categorica
initial_list_status	Estado inicial del préstamo. Puede ser W o F	préstamo	caracterizadora	categorica (binaria)
out_prncp	Cantidad restante del principal de la cantidad total pendiente por pagar	préstamo	balance	intervalo

Variable	Definición	Categoría	Subcategoría	Tipo de Variable
out_prncp_inv	Cantidad restante del principal de la cantidad en la que han invertido los inversores	préstamo	balance	intervalo
total_pymnt	Pagos recibidos a fecha corriente del total financiado	préstamo	pagos	intervalo
total_pymnt_inv	Pagos recibidos a fecha corriente del total en el que han invertido los inversores	préstamo	pagos	intervalo
total_rec_prncp	Parte del principal recibido a fecha de hoy	préstamo	pagos	intervalo
total_rec_int	Pago por tipo de interés recibido a fecha de hoy	préstamo	pagos	intervalo
total_rec_late_fee	Tasas por retraso recibidas a fecha de hoy	préstamo	recobro	intervalo
recoveries	Importe recuperado tras haber declarado "charge off"	préstamo	recobro	intervalo
collection_recovery_fee	Tasa de recuperación tras haber declarado "charge off"	préstamo	recobro	intervalo
last_pymnt_d	Último mes en el que se recibió un pago	préstamo	pagos	categorica
last_pymnt_amnt	Última cantidad total recibida con el último pago	préstamo	pagos	intervalo
next_pymnt_d	Próxima fecha de pago programada	préstamo	pagos	categorica
last_credit_pull_d	Último mes en el que LC tuvo que revisar el historial de crédito para este préstamo	solicitante	consultas	categorica
collections_12_mths_ex_med	Número de colecciones durante los últimos 12 meses excluyendo las médicas	préstamo	pagos	categorica
mths_since_last_major_derog	Meses transcurridos desde el peor rating de 90 días de duración	solicitante	calificación	categorica
policy_code	Número de poliza públicamente disponible=1; número de poliza de nuevos productos que no están públicamente disponibles=2.	solicitante	cuentas	categorica (binaria)
application_type	Tipo de aplicación: individual o conjunta (variable referencia*)	préstamo	caracterizadora	categorica (binaria)
annual_inc_joint	Combinación de la renta anual individual declarada por el pruncipal y el coprestatario en el momento de solicitud	solicitante	económica	intervalo
dti_joint	Ratio calculado usando los pagos mensuales de los co-prestatario sobre el total de las obligaciones de deuda.	solicitante	balance	intervalo
verification_status_joint	Estado de verificación de la renta: puede ser no verificado, verificado, verificada la fuente de ingreso	solicitante	social	categorica (binaria)
acc_now_delinq	Cuentas donde el prestatario es delincuente	solicitante	impago	categorica
tot_coll_amt	Cantidad máxima que ha debido alguna vez	solicitante	impago	intervalo
tot_cur_bal	Balance total de todas las cuentas	solicitante	balance	intervalo
open_acc_6m	Número de cuentas abiertas durante los últimos 6 meses	solicitante	cuentas	categorica
open_act_il	Número de cuentas de tipo "installment" activas	solicitante	cuentas	categorica
open_il_12m	Número de cuentas de tipo "installment" abiertas durante los últimos 12 meses	solicitante	cuentas	categorica
open_il_24m	Cuentas de tipo "installment" abiertas durante los últimos 24 meses	solicitante	cuentas	categorica
mths_since_rcnt_il	Meses transcurridos desde la apertura más reciente de cuentas de tipo "installment"	solicitante	cuentas	categorica
total_bal_il	Balance corriente total en cuentas de tipo installment	solicitante	balance	intervalo
il_util	Ratio: límite de crédito / balance nº de cuentas de tipo "installment"	solicitante	balance	intervalo
open_rv_12m	Cuentas de tipo "revolving" abiertas en los pasados 12 meses	solicitante	cuentas	categorica
open_rv_24m	Numero de cuentas tipo "revolving" abiertas durante los últimos 24 meses	solicitante	cuentas	categorica
max_bal_bc	Balance corriente máximo debido en todas las cuentas de tipo revolving	solicitante	balance	intervalo
all_util	Ratio: balance total del límite de crédito / balance de todas las cuentas	solicitante	balance	categorica
total_rev_hi_lim	Balance total de límite de crédito en cuentas "revolving"	solicitante	balance	intervalo
inq_fi	Nº de consultas financieras personales efectuadas para el individuo	solicitante	consultas	categorica
total_cu_tl	Número de cuentas financieras	solicitante	cuentas	categorica
inq_last_12m	Número de consultas crediticias realizadas en los últimos 12 meses	solicitante	consultas	categorica
acc_open_past_24mths	Número de cuentas abiertas durante los últimos 24 meses	solicitante	cuentas	categorica
avg_cur_bal	Balance medio corriente en todas las cuentas	solicitante	balance	intervalo

Variable	Definición	Categoría	Subcategoría	Tipo de Variable
bc_open_to_buy	Total disponible para utilizar en cuentas de tipo "revolving"	solicitante	balance	intervalo
bc_util	Ratio: de balance total/ límite de crédito de todas las tarjetas de crédito	solicitante	balance	intervalo
chargeoff_within_12_mths	Número de "charge-offs" dentro de los últimos 12 meses	solicitante	impago	categorica
delinq_amnt	Cantidad vencida de las cuentas en las que el prestatario está actualmente delinquiendo	solicitante	impago	intervalo
mo_sin_old_il_acct	Meses transcurridos desde que se abrió la primera cuenta de tipo "installment"	solicitante	cuentas	categorica
mo_sin_old_rev_tl_op	Meses transcurridos desde que se abrió la primera cuenta de tipo "revolving"	solicitante	cuentas	categorica
mo_sin_rcnt_rev_tl_op	Meses desde que se ha abierto la cuenta de tipo "revolving" más reciente	solicitante	cuentas	categorica
mo_sin_rcnt_tl	Meses desde que se ha abierto la cuenta más reciente	solicitante	cuentas	categorica
mort_acc	Número de hipotecas	solicitante	cuentas	categorica
mths_since_recent_bc	Meses desde la apertura más reciente de tarjetas de crédito	solicitante	cuentas	categorica
mths_since_recent_bc_dlq	Meses desde la delincuencia en tarjetas de crédito más reciente	solicitante	impago	categorica
mths_since_recent_inq	Meses transcurridos desde la consulta pública realizada más recientemente	solicitante	consultas	categorica
mths_since_recent_revol_delinq	Meses transcurridos desde la delincuencia más reciente	solicitante	impago	categorica
num_accts_ever_120_pd	Número de cuentas que alguna vez han estado más de 120 días vencidas	solicitante	impago	categorica
num_actv_bc_tl	Número de tarjetas de crédito actualmente activas	solicitante	cuentas	categorica
num_actv_rev_tl	Número de cuentas de tipo "revolving" actualmente activas	solicitante	cuentas	categorica
num_bc_sats	Número de tarjetas de crédito satisfactorias	solicitante	cuentas	categorica
num_bc_tl	Número de tarjetas de crédito	solicitante	cuentas	categorica
num_il_tl	Número de cuentas de tipo "installment"	solicitante	cuentas	categorica
num_op_rev_tl	Número de cuentas abiertas de tipo "revolving"	solicitante	cuentas	categorica
num_rev_accts	Número de cuentas de tipo "revolving"	solicitante	cuentas	categorica
num_rev_tl_bal_gt_0	Número de cuentas de tipo "revolving" con balance >0	solicitante	balance	categorica
num_sats	Número de cuentas satisfactorias	solicitante	cuentas	categorica
num_tl_120dpd_2m	Número de cuentas actualmente vencidas en más de 120 días (actualizado en los últimos dos meses)	solicitante	impago	categorica
num_tl_30dpd	Número de cuentas actualmente vencidas en más de 30 días (actualizado en los últimos dos meses)	solicitante	impago	categorica
num_tl_90g_dpd_24m	Número de cuentas vencias en 90 días o más durante los últimos 24 meses	solicitante	impago	categorica
num_tl_op_past_12m	Número de cuentas abiertas en los últimos 12 meses	solicitante	cuentas	categorica
pct_tl_nvr_dlq	Porcentaje de cuentas en las que nunca han sido delincuentes	solicitante	cuentas	categorica
percent_bc_gt_75	Porcentaje de todas las tarjetas de crédito con un endeudamiento total >75%	solicitante	balance	categorica
pub_rec_bankruptcies	Nº de veces que ha estado en bancarrota constatadas en el registro público	solicitante	económica	categorica
tax_liens	Nº de embargos fiscales constatados en el registro público	solicitante	económica	categorica
tot_hi_cred_lim	Balance total de límite de crédito en todas las cuentas	solicitante	balance	intervalo
total_bal_ex_mort	Balance total de crédito excluyendo el préstamo	solicitante	balance	intervalo
total_bc_limit	Balance total de límite de crédito en tarjetas de crédito	solicitante	balance	intervalo
total_il_high_credit_limit	Ratio: credito total en cuentas de tipo "installment"/límite de crédito	solicitante	balance	intervalo
revol_bal_joint	Suma del balance de crédito en todas las cuentas de tipo "revolving" de los co-prestatarios	solicitante	balance	categorica
sec_app_earliest_cr_line	Línea de crédito más antigua del prestatario secundario en el momento de solicitud	solicitante	cuentas	categorica
sec_app_inq_last_6mths	Consultas crediticias realizadas durante los últmpos 6 meses en el momento de solicitud del prestatario secundario	solicitante	consultas	categorica
sec_app_mort_acc	Número de hipotecas del prestatario secundario en el momento de solicitud	solicitante	cuentas	categorica

Variable	Definición	Categoría	Subcategoría	Tipo de Variable
sec_app_open_acc	Número de cuentas abiertas en el momento de solicitud del prestatario secundario	solicitante	cuentas	categorica
sec_app_revol_util	Ratio: balance corriente / límite de endeudamiento en cuentas de tipo "revolving" del prestatario secundario	solicitante	balance	categorica
sec_app_open_act_il	Número de cuentas actualmente activas de tipo "installment" en el momento de solicitud del prestatario secundario	solicitante	cuentas	categorica
sec_app_num_rev_accts	Número de cuentas de tipo revolving en el momento de solicitud del prestatario secundario	solicitante	cuentas	categorica
sec_app_chargeoff_within_12_mths	Número de "charge-offs" dentro de los últimos 12 meses del prestatario secundario	solicitante	impago	categorica
sec_app_collections_12_mths_ex_med	Número de colecciones de pagos realizadas dentro de los últimos 12 meses excluyendo las de tipo médico de prestatario secundario	préstamo	pagos	categorica
sec_app_mths_since_last_major_derog	Meses transcurridos desde el peor rating de 90 días de duración del prestatario secundario	solicitante	calificación	categorica
hardship_flag	Indicador de si el prestatario ha tenido que recurrir a plan de dificultad	préstamo	reestructuración	categorica (binaria)
hardship_type	Descripción del plan de dificultad ofrecido	préstamo	reestructuración	categorica
hardship_reason	Razón por la que se le ofreció el plan de dificultad	préstamo	reestructuración	categorica
hardship_status	Estado del plan de dificultad: active, pending, completed, cancelled o broken	préstamo	reestructuración	categorica
deferral_term	Cantidad de meses que se espera que el prestatario pague menos de lo estipulado en el contrato debido a que es partícipe de un plan de dificultad	préstamo	reestructuración	categorica
hardship_amount	Pago de intereses que el prestatario ha acordado realizar durante el plan de dificultad	préstamo	reestructuración	categorica
hardship_start_date	Fecha de inicio del periodo del plan de dificultad	préstamo	reestructuración	categorica
hardship_end_date	Fecha de fin del plan de dificultad	préstamo	reestructuración	categorica
payment_plan_start_date	Primer día de finalización del plan de dificultad, calculado a partir del inicio y de la duración estipulada del mismo	préstamo	reestructuración	categorica
hardship_length	Número de meses estipulados de reducción de cantidad a pagar por el prestatario debido a la participación en un plan de dificultad	préstamo	reestructuración	categorica
hardship_dpd	Número de días de vencimiento de la cuenta cuando comienza el plan de dificultad	préstamo	reestructuración	categorica
hardship_loan_status	Estado del préstamo en el momento que comienza el plan de dificultad	préstamo	reestructuración	categorica
orig_projected_additional_accrued_interest	El monto de interés adicional respecto al original que se acumulará para el plan de pago por dificultad a partir de la fecha de su inicio	préstamo	reestructuración	intervalo
hardship_payoff_balance_amount	Balance pendiente por pagar en fecha de inicio del plan de dificultad	préstamo	reestructuración	intervalo
hardship_last_payment_amount	Última cantidad que fue pagada antes de comenzar el plan de dificultad	préstamo	reestructuración	intervalo
debt_settlement_flag	Indicador de si el prestatario que ha incurrido en "charge-off" está trabajando con una empresa externa de liquidación de deuda	préstamo	liquidación	categorica
debt_settlement_flag_d	Fecha más reciente de actualización de la variable debt_settlement_flag	préstamo	liquidación	categorica
settlement_date	Fecha en la que el prestatario acepta el plan de liquidación de deuda	préstamo	liquidación	categorica
settlement_amount	Cantidad del préstamo que ha acordado devolver con el plan de liquidación de deuda	préstamo	liquidación	intervalo
settlement_percentage	Cantidad devuelta como porcentaje del balance de la cantidad total pendiente por devolver	préstamo	liquidación	intervalo
settlement_term	Número de meses que el prestatario estará con el plan de liquidación de deuda	préstamo	liquidación	categórica
Debt_settlement_status	Estado del plan de liquidación de deuda	préstamo	liquidación	categórica

Anexo II - Matriz Análisis Factorial

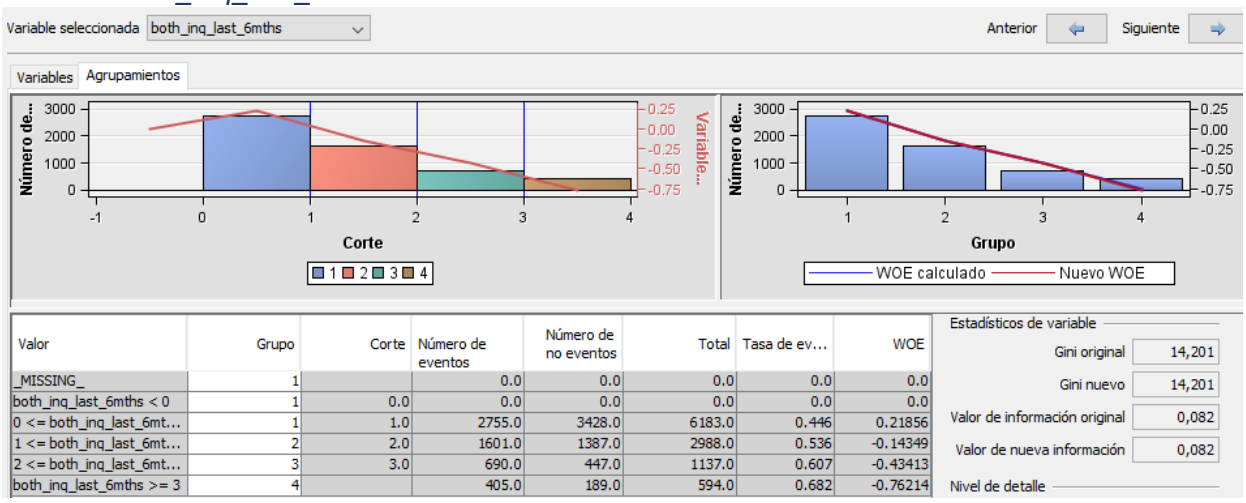
	Componente													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
num_op_rev_tl	.908	.012	-.045	-.088	.262	.062	-.009	-.063	-.021	.037	-.057	.021	-.003	-.064
num_sats	.837	.352	.033	-.041	.204	.017	.095	-.076	-.009	.022	.056	-.001	.015	.062
num_rev_tl_bal_gt_0	.810	.003	-.020	.238	.159	.071	-.031	-.165	.011	.031	-.055	.000	-.018	-.079
num_bc_sats	.804	-.011	-.028	-.175	.137	.278	.016	-.168	-.005	.058	-.021	.010	-.049	-.150
both_num_rev_accts	.764	.006	.034	-.042	.141	-.005	-.038	.371	-.019	.102	.123	.114	.060	.216
both_open_acc	.760	.284	.066	.029	.105	-.067	.053	-.024	-.071	.101	.251	.101	.062	.350
num_bc_tl	.744	-.013	-.002	-.190	.147	.236	.003	.313	.051	.055	-.034	.023	-.039	-.132
total_acc	.629	.417	.121	-.057	.162	.027	.150	.482	.050	.037	.050	.011	.040	.006
total_bal_il	.059	.921	.181	.015	.039	.012	.089	.030	.022	.068	.139	.059	-.006	-.097
nom_total_bal_il	-.041	.883	.039	-.024	.042	-.062	.094	-.032	.016	-.051	-.262	.023	-.021	.004
total_il_high_credit_limit	.101	.879	.196	.024	-.002	.046	.066	.051	.003	.091	.208	.049	.011	-.093
nom_tot_bal_ex_mort	.060	.860	.082	.085	.029	.133	.007	-.026	-.023	-.048	-.365	.004	-.026	.021
total_bal_ex_mort	.190	.834	.245	.123	.011	.227	.043	.036	-.010	.121	.178	.049	-.012	-.088
num_il_tl	.125	.684	.058	.027	.025	-.050	.249	.318	.027	.009	.104	-.023	.037	.043
both_open_act_il	.133	.650	-.001	.099	-.066	-.120	.180	-.035	-.033	.019	.294	.023	.061	.390
il_util	-.028	.506	.036	.034	.069	-.077	.424	-.096	.074	.020	-.077	.014	-.040	.083
dti_total	.274	.436	-.041	.232	-.091	.154	.145	.007	-.114	.004	-.382	.026	.043	.293
avg_cur_bal	-.162	.072	.910	.071	-.045	.068	.014	.093	.034	.106	.146	.020	.004	-.043
nom_tot_cur_bal	.067	.237	.887	.018	.043	.013	.031	.020	.011	-.034	-.246	-.011	.001	.047
tot_cur_bal	.157	.239	.875	.052	.021	.110	.055	.042	.018	.108	.238	.026	.008	-.038
nom_avg_cur_bal	-.274	.041	.852	.042	-.039	-.008	-.013	.071	.024	-.030	-.281	-.011	-.008	.020
tot_hi_cred_lim	.230	.223	.842	-.028	.011	.175	.052	.056	.007	.114	.269	.021	.010	-.037
both_mort_acc	.173	-.033	.610	.032	-.056	.034	.097	.288	.007	.105	.230	.059	.058	.235
bc_util	-.014	.026	.038	.924	-.095	.108	-.060	.002	.015	.012	-.005	-.062	-.009	-.072
both_revol_util	-.034	.026	.069	.897	-.135	.157	-.055	-.018	.020	.051	.051	-.046	-.003	-.008
percent_bc_gt_75	.014	.016	.045	.827	-.100	.058	-.056	.028	-.006	-.011	.032	-.061	-.020	-.077
all_util	-.149	.355	.036	.690	-.051	-.026	.235	-.095	.095	.027	-.036	-.018	-.032	.060
bc_open_to_buy	.374	.012	.074	-.645	.011	.457	.026	.014	-.087	.039	.132	-.015	-.057	-.069
open_rv_12m	.215	-.049	-.069	-.085	.878	-.029	-.108	-.043	.045	-.022	-.026	.130	-.004	-.021
num_tl_op_past_12m	.152	.083	.043	-.105	.855	-.019	.315	-.001	.021	-.018	.023	.171	.003	.023
open_rv_24m	.396	-.051	-.100	-.084	.777	-.101	-.062	-.076	.032	-.010	-.079	.058	-.046	-.076
acc_open_past_24mths	.349	.126	.037	-.084	.751	-.104	.354	-.009	-.004	.004	-.015	.055	-.038	-.036
open_acc_6m	.058	.055	.032	-.118	.739	.018	.133	.011	.015	-.033	.061	.239	.042	.086
max_bal_bc	.115	.083	.172	.313	-.081	.766	-.042	.101	-.109	.177	.134	-.043	-.008	-.025
nom_max_bal_bc	-.018	-.009	-.023	.291	-.061	.698	-.105	.067	-.150	-.006	-.396	-.085	-.009	.089
total_bc_limit	.471	.044	.131	-.360	-.033	.689	-.005	.013	-.121	.120	.166	-.034	-.059	-.076
total_rev_hi_lim	.573	.067	.174	-.289	.016	.610	-.013	.045	-.135	.131	.160	-.013	-.038	-.012
nom_revol_bal	.369	-.002	.064	.378	-.083	.541	-.117	.022	-.177	.091	-.240	.005	.018	.335
revol_bal_alltotal	.415	.077	.205	.348	-.099	.498	-.049	.057	-.133	.237	.264	.052	.024	.249
open_il_24m	.027	.363	.044	-.017	.186	-.028	.794	.094	-.049	.005	.084	.001	.005	.046
open_il_12m	-.043	.263	.038	-.054	.240	.022	.779	.056	-.024	-.014	.076	.100	.019	.076
inq_fi	.090	.071	.086	-.003	-.020	-.118	.600	.017	.043	.011	-.084	.362	-.036	-.220
pct_all_sats	.136	-.110	-.160	.017	.068	-.036	-.098	-.852	-.124	-.052	.007	-.010	-.038	.053
pct_bc_sats	.028	-.004	-.063	.031	-.009	.039	.024	-.814	-.108	-.018	.021	-.019	-.013	-.024
mths_earliest_cr_line	.247	-.007	.140	.032	-.105	.206	.001	.448	.136	.000	.085	-.016	.032	-.018
pct_tl_nvr_dlq	.107	.040	-.014	-.048	.020	.093	.040	-.054	-.862	.023	-.017	-.029	-.122	-.021
delinq_account	.056	.021	.049	.045	.020	-.097	.028	.125	.767	-.009	.035	.020	.121	-.010
num_accts_ever_120_pd	-.006	.043	-.011	.016	.102	-.080	.012	.150	.742	-.003	-.028	.023	.027	.053
pub_rec_bankruptcies	.047	-.016	-.036	.078	.076	-.319	-.006	.166	-.320	.042	-.115	.083	.065	-.318
loan_amnt	.125	.055	.096	-.005	-.029	.186	-.030	.034	-.018	.920	.182	-.003	-.013	-.062
installment	.119	.057	.057	.055	-.005	.209	-.035	.008	.005	.828	.211	.007	.000	-.120
term	.040	.027	.086	.006	-.027	-.094	.060	.041	-.031	.602	-.151	.023	.000	.192
annual_inc_both	.195	.121	.278	.063	-.037	.155	.067	.080	.035	.281	.773	.072	.021	-.134
mths_since_recent_inq	-.013	-.049	-.026	.086	-.171	.011	-.090	-.009	-.040	-.010	.011	-.802	.012	.075
both_inq_last_6mths	.055	.021	-.038	-.036	.208	-.026	-.043	.026	-.023	-.006	.126	.762	.047	.195
inq_last_12m	.066	.069	.086	-.053	.213	-.066	.362	-.005	.055	.039	-.059	.697	-.014	-.169
num_tl_90g_venc24	-.013	-.002	.009	-.012	-.011	-.001	.005	.005	.242	.003	-.043	.011	.791	-.122
delinq_2yrs	.049	.006	.060	.003	-.077	-.041	-.012	.030	.408	-.016	.018	.029	.692	-.122
both_chargeoff_12m	-.017	-.002	-.008	-.011	.032	-.013	-.009	.034	-.100	-.002	.028	-.008	.527	.093
tax_liens	.013	.005	-.016	.030	-.013	-.020	.011	-.003	-.026	-.008	.052	.008	.021	-.288

Anexo III - Detección datos atípicos

	Método 1			R. Intercuartílico		LIMITES		Atípicos encontrados
variable	Método	L. INF	L. SUP	L. INF	L. SUP	L. INF	L. SUP	
acc_open_past_24mths	MADS	.	22	.	19	.	22	141
all_util	STDDEV	.	118.04	.	.	.	.	35
annual_inc_both	MADS	.	316000	.	271000	.	316000	1129
both_mort_acc	MADS	.	10	.	9	.	10	639
delinq_2yrs	PERCENTS	.	4	.	1	.	4	611
dti_total	STDDEV	.	43.74	.	.	.	.	0
il_util	STDDEV	.	156.6	.	213	.	213	20
inq_fi	MADS	.	10	.	9	.	10	217
installment	STDDEV	.	1321.72	.	.	.	.	0
loan_amnt	STDDEV	.	46297.71	.	.	.	.	0
mths_earliest_cr_line	STDDEV	.	473.93	.	553	.	553	467
nom_revol_bal	MADS	.	100.14	.	86.84	.	100.14	675
nom_tot_bal_ex_mort	MADS	.	267.54	.	231.35	.	267.54	1737
nom_tot_cur_bal	MADS	.	771.35	.	854.45	.	854.45	129
nom_total_bal_il	MADS	.	222.52	.	194.41	.	222.52	2212
num_rev_tl_bal_gt_0	MADS	.	23	.	20	.	23	155
open_il_24m	MADS	.	10	.	9	.	10	104
open_rv_12m	MADS	.	10	.	9	.	10	166
pct_all_sats	STDDEV	.	112.02	.	.	.	.	0
pct_tl_nvr_dlq	MADS	.	.	.	9.1	.	.	0
tot_hi_cred_lim	MADS	.	825562	.	901024	.	901024	781
total_acc	MADS	.	83	.	75	.	83	111
total_bc_limit	MADS	.	116200	.	107100	.	116200	1421
total_rev_hi_lim	MADS	.	157100	.	141900	.	157100	1629
both_chargeoff_12m	*PERCENTS*	*.*	*1*	*.*	*1*	*.*	*1*	*0*
pub_rec_bankruptcies	*PERCENTS*	*.*	*1*	*.*	*1*	*.*	*1*	*0*
both_inq_last_6mths	*PERCENTS*	*.*	*1*	*.*	*1*	*.*	*1*	*0*
tax_liens	*PERCENTS*	*.*	*1*	*.*	*1*	*.*	*1*	*0*

Anexo IV - Tramificación WOE de las variables del estudio

Variable both_inq_last_6mths



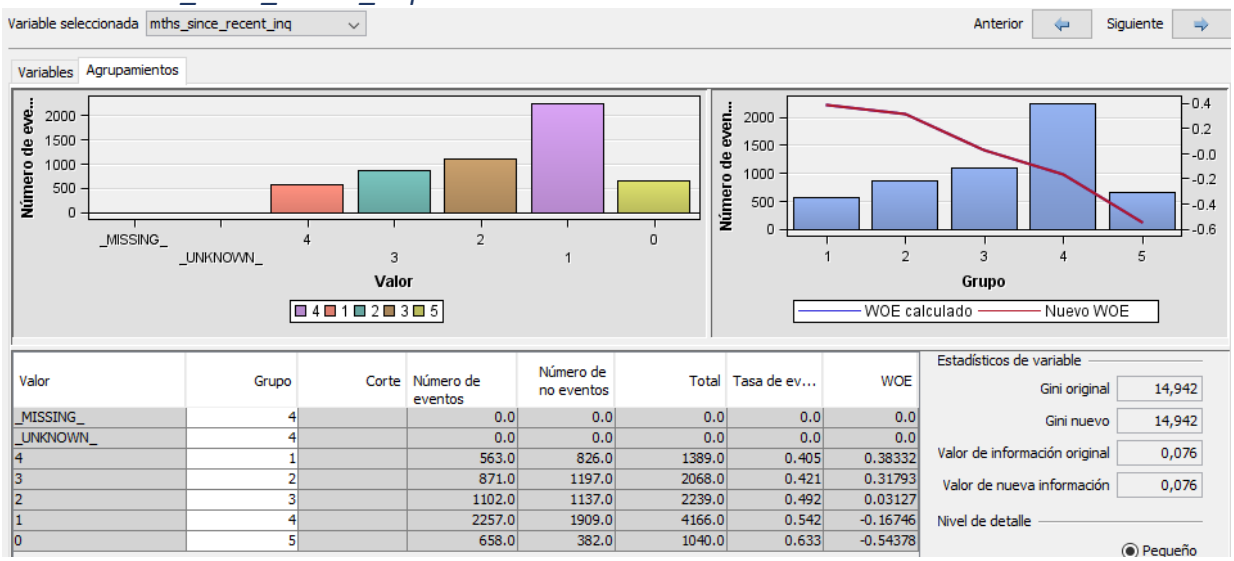
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
both_inq_last_6mths < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= both_inq_last_6mt...	1	1.0	2755.0	3428.0	6183.0	0.446	0.21856
1 <= both_inq_last_6mt...	2	2.0	1601.0	1387.0	2988.0	0.536	-0.14349
2 <= both_inq_last_6mt...	3	3.0	690.0	447.0	1137.0	0.607	-0.43413
both_inq_last_6mths >= 3	4		405.0	189.0	594.0	0.682	-0.76214

Estadísticos de variable

Gini original	14,201
Gini nuevo	14,201
Valor de información original	0,082
Valor de nueva información	0,082

Nivel de detalle

Variable mths_since_recent_inq



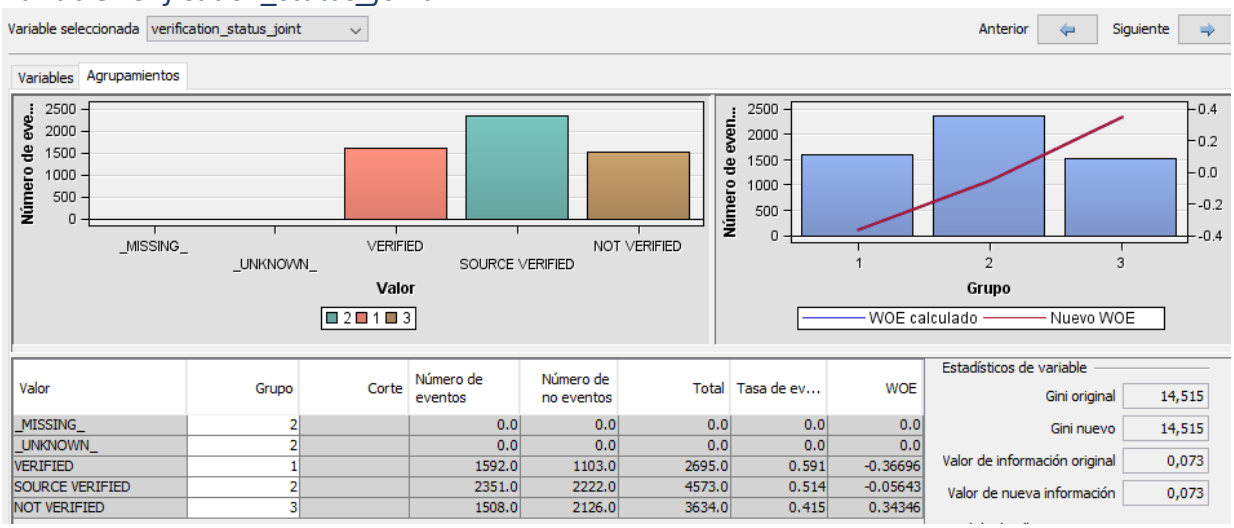
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
UNKNOWN	4		0.0	0.0	0.0	0.0	0.0
4	1		563.0	826.0	1389.0	0.405	0.38332
3	2		871.0	1197.0	2068.0	0.421	0.31793
2	3		1102.0	1137.0	2239.0	0.492	0.03127
1	4		2257.0	1909.0	4166.0	0.542	-0.16746
0	5		658.0	382.0	1040.0	0.633	-0.54378

Estadísticos de variable

Gini original	14,942
Gini nuevo	14,942
Valor de información original	0,076
Valor de nueva información	0,076

Nivel de detalle: Pequeño

Variable verification_status_joint

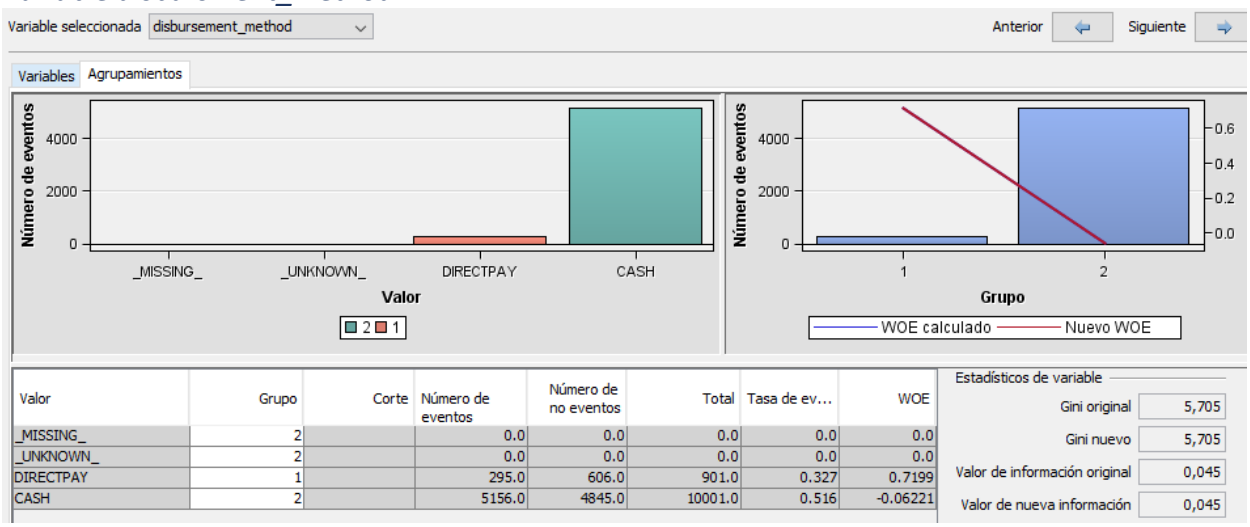


Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	2		0.0	0.0	0.0	0.0	0.0
UNKNOWN	2		0.0	0.0	0.0	0.0	0.0
VERIFIED	1		1592.0	1103.0	2695.0	0.591	-0.36696
SOURCE VERIFIED	2		2351.0	2222.0	4573.0	0.514	-0.05643
NOT VERIFIED	3		1508.0	2126.0	3634.0	0.415	0.34346

Estadísticos de variable

Gini original	14,515
Gini nuevo	14,515
Valor de información original	0,073
Valor de nueva información	0,073

Variable disbursment_method

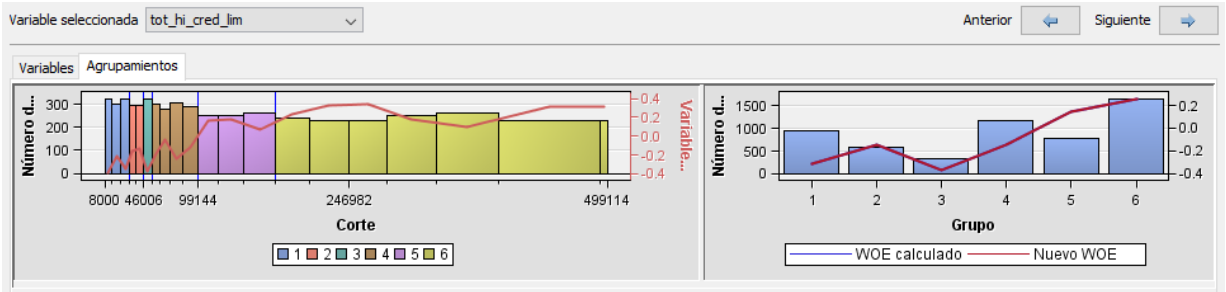


Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	2		0.0	0.0	0.0	0.0	0.0
UNKNOWN	2		0.0	0.0	0.0	0.0	0.0
DIRECTPAY	1		295.0	606.0	901.0	0.327	0.7199
CASH	2		5156.0	4845.0	10001.0	0.516	-0.06221

Estadísticos de variable

Gini original	5,705
Gini nuevo	5,705
Valor de información original	0,045
Valor de nueva información	0,045

Variable tot_hi_cred_lim



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	6		0.0	0.0	0.0	0.0	0.0
tot_hi_cred_lim < 15100	1	15100.0	324.0	220.0	544.0	0.596	-0.38712
15100 <= tot_hi_cred_li...	1	24103.0	302.0	244.0	546.0	0.553	-0.21326
24103 <= tot_hi_cred_li...	1	31700.0	319.0	223.0	542.0	0.589	-0.35802
31700 <= tot_hi_cred_li...	2	38800.0	295.0	250.0	545.0	0.541	-0.16551
38800 <= tot_hi_cred_li...	2	46006.0	293.0	255.0	548.0	0.535	-0.13891
46006 <= tot_hi_cred_li...	3	53975.0	323.0	222.0	545.0	0.593	-0.37497
53975 <= tot_hi_cred_li...	4	62324.0	298.0	247.0	545.0	0.547	-0.18771
62324 <= tot_hi_cred_li...	4	72102.0	278.0	267.0	545.0	0.51	-0.04037
72102 <= tot_hi_cred_li...	4	84393.0	305.0	240.0	545.0	0.56	-0.23967
84393 <= tot_hi_cred_li...	4	99144.0	289.0	256.0	545.0	0.53	-0.12125
99144 <= tot_hi_cred_li...	5	119198.0	250.0	296.0	546.0	0.458	0.1689
119198 <= tot_hi_cred_...	5	144401.0	249.0	296.0	545.0	0.457	0.17291
144401 <= tot_hi_cred_...	5	174822.0	263.0	282.0	545.0	0.483	0.06975
174822 <= tot_hi_cred_...	6	207680.0	241.0	304.0	545.0	0.442	0.23223
207680 <= tot_hi_cred_...	6	246982.0	229.0	316.0	545.0	0.42	0.32202
246982 <= tot_hi_cred_...	6	284093.0	226.0	319.0	545.0	0.415	0.34466
284093 <= tot_hi_cred_...	6	332000.0	248.0	297.0	545.0	0.455	0.1803
332000 <= tot_hi_cred_...	6	392652.0	259.0	286.0	545.0	0.475	0.09916
392652 <= tot_hi_cred_...	6	492014.0	230.0	315.0	545.0	0.422	0.31449
tot_hi_cred_lim >= 492014	6		230.0	316.0	546.0	0.421	0.31766

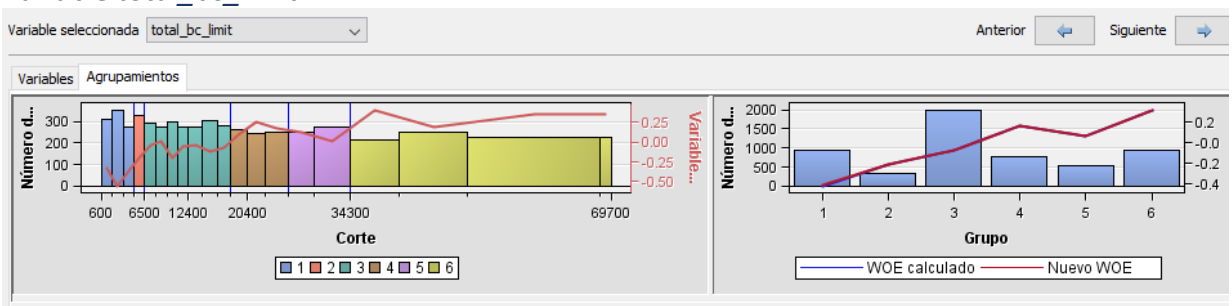
Estadísticos de variable

Gini original	12,812
Gini nuevo	12,812
Valor de información original	0,055
Valor de nueva información	0,055

Nivel de detalle: (•) Pequeño ( ) Grande

Rol Variable: Predeterminado — Aplicar

Variable total_bc_limit



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	3		0.0	0.0	0.0	0.0	0.0
total_bc_limit < 2000	1	2000.0	309.0	220.0	529.0	0.584	-0.33971
2000 <= total_bc_limit <...	1	3600.0	355.0	204.0	559.0	0.635	-0.554
3600 <= total_bc_limit <...	1	5000.0	273.0	190.0	463.0	0.59	-0.36245
5000 <= total_bc_limit <...	2	6500.0	327.0	265.0	592.0	0.552	-0.21023
6500 <= total_bc_limit <...	3	8000.0	290.0	277.0	567.0	0.511	-0.04586
8000 <= total_bc_limit <...	3	9500.0	272.0	276.0	548.0	0.496	0.0146
9500 <= total_bc_limit <...	3	10900.0	297.0	242.0	539.0	0.551	-0.20479
10900 <= total_bc_limit ...	3	12400.0	276.0	262.0	538.0	0.513	-0.05206
12400 <= total_bc_limit ...	3	14300.0	274.0	264.0	538.0	0.509	-0.03718
14300 <= total_bc_limit ...	3	16400.0	305.0	270.0	575.0	0.53	-0.12189
16400 <= total_bc_limit ...	3	18100.0	277.0	258.0	535.0	0.518	-0.07106
18100 <= total_bc_limit ...	4	20400.0	261.0	286.0	547.0	0.477	0.09147
20400 <= total_bc_limit ...	4	22900.0	242.0	312.0	554.0	0.437	0.25407
22900 <= total_bc_limit ...	4	26000.0	249.0	295.0	544.0	0.458	0.16952
26000 <= total_bc_limit ...	5	29500.0	251.0	283.0	534.0	0.47	0.11999
29500 <= total_bc_limit ...	5	34300.0	276.0	279.0	555.0	0.497	0.01081
34300 <= total_bc_limit ...	6	41000.0	215.0	324.0	539.0	0.399	0.41011
41000 <= total_bc_limit ...	6	50300.0	251.0	301.0	552.0	0.455	0.18166
50300 <= total_bc_limit ...	6	68300.0	226.0	322.0	548.0	0.412	0.35402
total_bc_limit >= 68300	6		225.0	321.0	546.0	0.412	0.35534

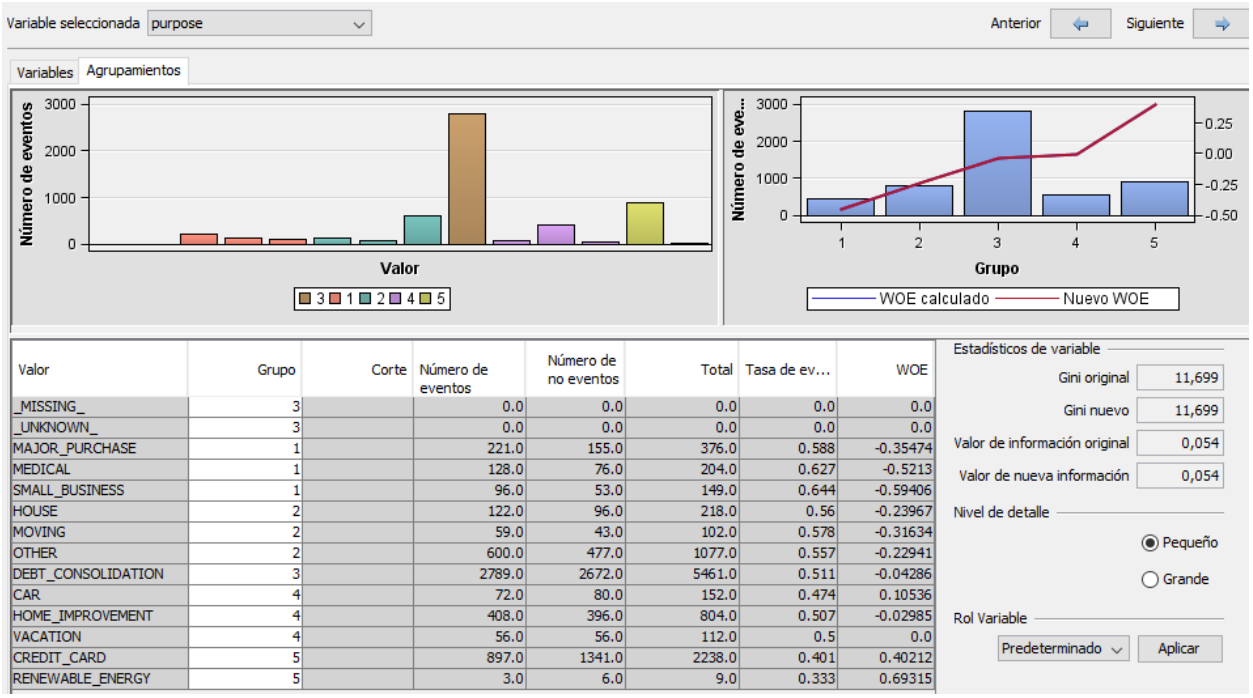
Estadísticos de variable

Gini original	12,933
Gini nuevo	12,933
Valor de información original	0,055
Valor de nueva información	0,055

Nivel de detalle: (•) Pequeño ( ) Grande

Rol Variable: Predeterminado — Aplicar

Variable purpose



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	3		0.0	0.0	0.0	0.0	0.0
UNKNOWN	3		0.0	0.0	0.0	0.0	0.0
MAJOR_PURCHASE	1		221.0	155.0	376.0	0.588	-0.35474
MEDICAL	1		128.0	76.0	204.0	0.627	-0.5213
SMALL_BUSINESS	1		96.0	53.0	149.0	0.644	-0.59406
HOUSE	2		122.0	96.0	218.0	0.56	-0.23967
MOVING	2		59.0	43.0	102.0	0.578	-0.31634
OTHER	2		600.0	477.0	1077.0	0.557	-0.22941
DEBT_CONSOLIDATION	3		2789.0	2672.0	5461.0	0.511	-0.04286
CAR	4		72.0	80.0	152.0	0.474	0.10536
HOME_IMPROVEMENT	4		408.0	396.0	804.0	0.507	-0.02985
VACATION	4		56.0	56.0	112.0	0.5	0.0
CREDIT_CARD	5		897.0	1341.0	2238.0	0.401	0.40212
RENEWABLE_ENERGY	5		3.0	6.0	9.0	0.333	0.69315

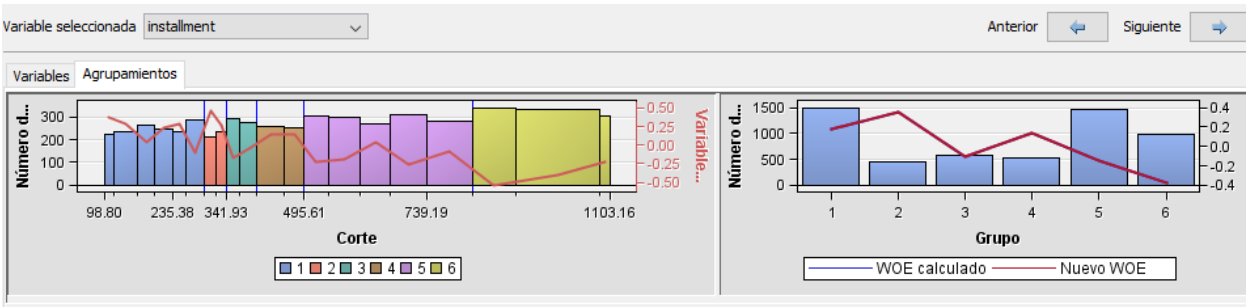
Estadísticos de variable

Gini original	11,699
Gini nuevo	11,699
Valor de información original	0,054
Valor de nueva información	0,054

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Variable installment



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
installment < 119.54	1	119.54	223.0	322.0	545.0	0.409	0.36738
119.54 <= installment <...	1	166.03	233.0	309.0	542.0	0.43	0.2823
166.03 <= installment <...	1	199.23	265.0	277.0	542.0	0.489	0.04429
199.23 <= installment <...	1	235.38	245.0	306.0	551.0	0.445	0.22233
235.38 <= installment <...	1	261.86	234.0	311.0	545.0	0.429	0.28447
261.86 <= installment <...	1	298.29	286.0	258.0	544.0	0.526	-0.10303
298.29 <= installment <...	2	321.19	212.0	334.0	546.0	0.388	0.45455
321.19 <= installment <...	2	341.93	235.0	306.0	541.0	0.434	0.264
341.93 <= installment <...	3	368.69	294.0	249.0	543.0	0.541	-0.16613
368.69 <= installment <...	3	402.08	276.0	264.0	540.0	0.511	-0.04445
402.08 <= installment <...	4	457.36	259.0	296.0	555.0	0.467	0.13353
457.36 <= installment <...	4	495.61	251.0	290.0	541.0	0.464	0.14443
495.61 <= installment <...	5	544.19	306.0	243.0	549.0	0.557	-0.23052
544.19 <= installment <...	5	607.97	298.0	247.0	545.0	0.547	-0.18771
607.97 <= installment <...	5	667.79	267.0	279.0	546.0	0.489	0.04396
667.79 <= installment <...	5	739.19	307.0	237.0	544.0	0.564	-0.25879
739.19 <= installment <...	5	830.12	283.0	259.0	542.0	0.522	-0.08862
830.12 <= installment <...	6	916.03	341.0	200.0	541.0	0.63	-0.53357
916.03 <= installment <...	6	1082.42	333.0	221.0	554.0	0.601	-0.40998
installment >= 1082.42	6		303.0	243.0	546.0	0.555	-0.22067

Estadísticos de variable

Gini original	12,652
Gini nuevo	12,652
Valor de información original	0,053
Valor de nueva información	0,053

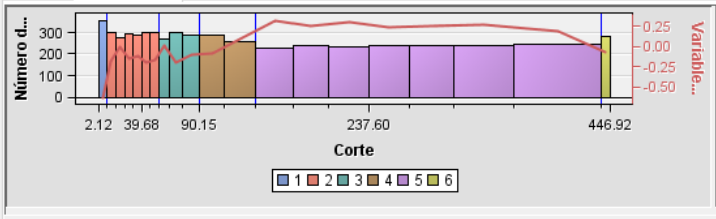
Nivel de detalle: Pequeño / Grande

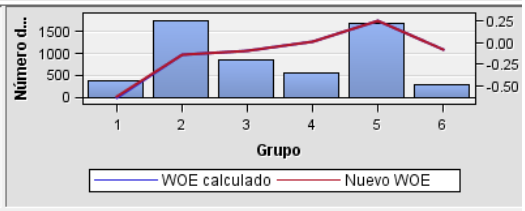
Rol Variable: Predeterminado — Aplicar

Variable nom_tot_cur_bal

Variable seleccionada: nom_tot_cur_bal — Anterior — Siguiente

Variables | Agrupamientos





Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	5		0.0	0.0	0.0	0.0	0.0
nom_tot_cur_bal < 9.09	1	9.09	356.0	190.0	546.0	0.652	-0.62791
9.09 <= nom_tot_cur_b...	2	17.04	296.0	248.0	544.0	0.544	-0.17693
17.04 <= nom_tot_cur_...	2	25.04	274.0	271.0	545.0	0.503	-0.01101
25.04 <= nom_tot_cur_...	2	32.71	294.0	252.0	546.0	0.538	-0.15415
32.71 <= nom_tot_cur_...	2	39.68	289.0	255.0	544.0	0.531	-0.12516
39.68 <= nom_tot_cur_...	2	46.83	299.0	245.0	544.0	0.55	-0.19919
46.83 <= nom_tot_cur_...	2	55.13	296.0	250.0	546.0	0.542	-0.1689
55.13 <= nom_tot_cur_...	3	64.22	271.0	274.0	545.0	0.497	0.01101
64.22 <= nom_tot_cur_...	3	75.54	301.0	245.0	546.0	0.551	-0.20585
75.54 <= nom_tot_cur_...	3	90.15	287.0	258.0	545.0	0.527	-0.10652
90.15 <= nom_tot_cur_...	4	111.59	284.0	262.0	546.0	0.52	-0.08063
111.59 <= nom_tot_cur...	4	139.1	259.0	285.0	544.0	0.476	0.09566
139.1 <= nom_tot_cur_...	5	171.6	229.0	317.0	546.0	0.419	0.32518
171.6 <= nom_tot_cur_...	5	202.37	239.0	307.0	546.0	0.438	0.25038
202.37 <= nom_tot_cur...	5	237.6	232.0	311.0	543.0	0.427	0.29306
237.6 <= nom_tot_cur_...	5	272.37	241.0	305.0	546.0	0.441	0.23551
272.37 <= nom_tot_cur...	5	311.31	238.0	307.0	545.0	0.437	0.25458
311.31 <= nom_tot_cur...	5	364.07	236.0	309.0	545.0	0.433	0.26951
364.07 <= nom_tot_cur...	5	439.95	247.0	297.0	544.0	0.454	0.18434
nom_tot_cur_bal >= 43...	6		283.0	263.0	546.0	0.518	-0.07329

Estadísticos de variable

Gini original	11,469
Gini nuevo	11,469
Valor de información original	0,050
Valor de nueva información	0,050

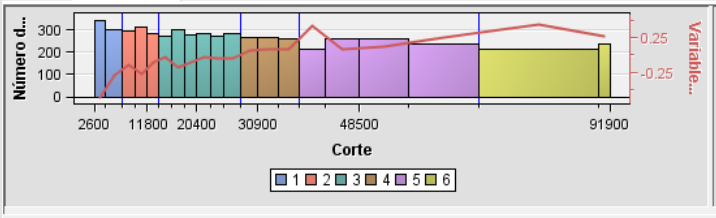
Nivel de detalle: (•) Pequeño ( ) Grande

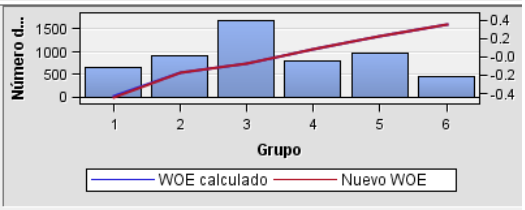
Rol Variable: Predeterminado — Aplicar

Variable total_rev_hi_lim

Variable seleccionada: total_rev_hi_lim — Anterior — Siguiente

Variables | Agrupamientos





Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	3		0.0	0.0	0.0	0.0	0.0
total_rev_hi_lim < 4600	1	4600.0	345.0	190.0	535.0	0.645	-0.59652
4600 <= total_rev_hi_li...	1	7400.0	304.0	228.0	532.0	0.571	-0.28768
7400 <= total_rev_hi_li...	2	9800.0	298.0	261.0	559.0	0.533	-0.13257
9800 <= total_rev_hi_li...	2	11800.0	311.0	238.0	549.0	0.566	-0.26752
11800 <= total_rev_hi_li...	2	13800.0	283.0	253.0	536.0	0.528	-0.11206
13800 <= total_rev_hi_li...	3	16000.0	272.0	265.0	537.0	0.507	-0.02607
16000 <= total_rev_hi_li...	3	18200.0	304.0	256.0	560.0	0.543	-0.17185
18200 <= total_rev_hi_li...	3	20400.0	280.0	254.0	534.0	0.524	-0.09746
20400 <= total_rev_hi_li...	3	22800.0	284.0	276.0	560.0	0.507	-0.02857
22800 <= total_rev_hi_li...	3	25100.0	271.0	257.0	528.0	0.513	-0.05304
25100 <= total_rev_hi_li...	3	28000.0	285.0	273.0	558.0	0.511	-0.04302
28000 <= total_rev_hi_li...	4	30900.0	267.0	283.0	550.0	0.485	0.0582
30900 <= total_rev_hi_li...	4	34500.0	264.0	284.0	548.0	0.482	0.07303
34500 <= total_rev_hi_li...	4	38100.0	260.0	284.0	544.0	0.478	0.08829
38100 <= total_rev_hi_li...	5	42600.0	214.0	327.0	541.0	0.396	0.42398
42600 <= total_rev_hi_li...	5	48500.0	262.0	284.0	546.0	0.48	0.08063
48500 <= total_rev_hi_li...	5	57000.0	258.0	291.0	549.0	0.47	0.12036
57000 <= total_rev_hi_li...	5	69300.0	239.0	305.0	544.0	0.439	0.24385
69300 <= total_rev_hi_li...	6	89900.0	214.0	332.0	546.0	0.392	0.43916
total_rev_hi_lim >= 89900	6		236.0	310.0	546.0	0.432	0.27274

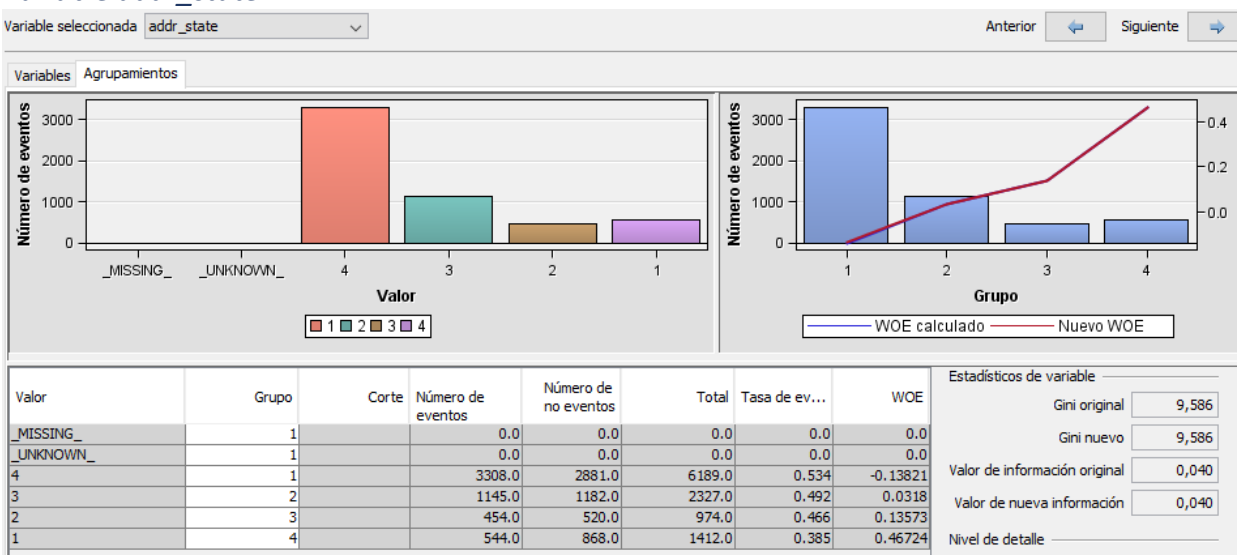
Estadísticos de variable

Gini original	11,946
Gini nuevo	11,946
Valor de información original	0,047
Valor de nueva información	0,047

Nivel de detalle: (•) Pequeño ( ) Grande

Rol Variable: Predeterminado — Aplicar

Variable addr_state

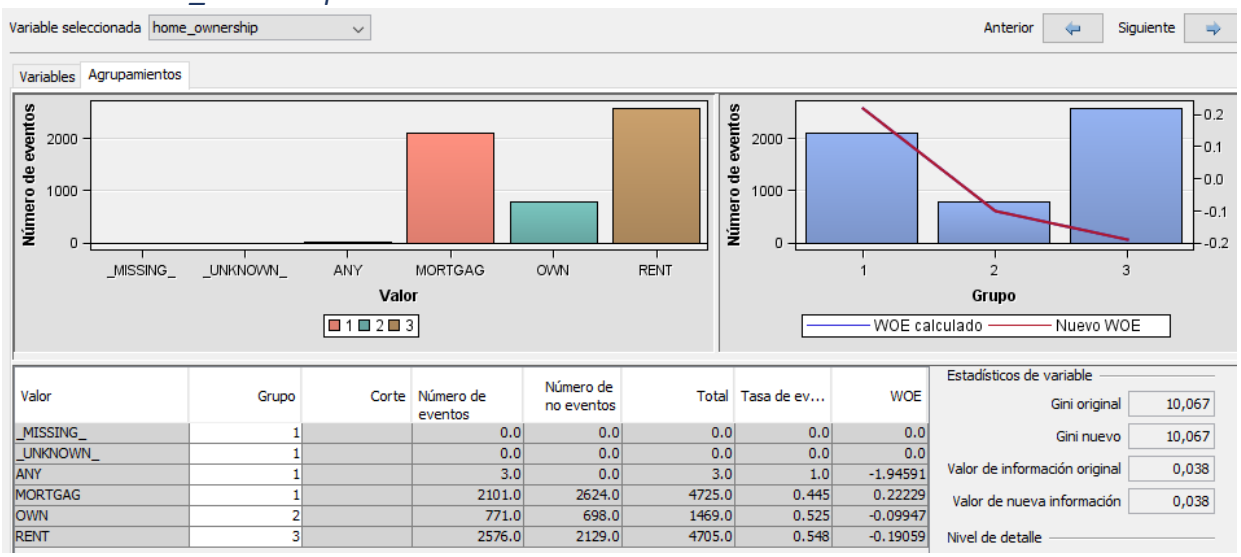


Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
UNKNOWN	1		0.0	0.0	0.0	0.0	0.0
4	1		3308.0	2881.0	6189.0	0.534	-0.13821
3	2		1145.0	1182.0	2327.0	0.492	0.0318
2	3		454.0	520.0	974.0	0.466	0.13573
1	4		544.0	868.0	1412.0	0.385	0.46724

Estadísticos de variable

Gini original	9,586
Gini nuevo	9,586
Valor de información original	0,040
Valor de nueva información	0,040
Nivel de detalle	

Variable home_ownership

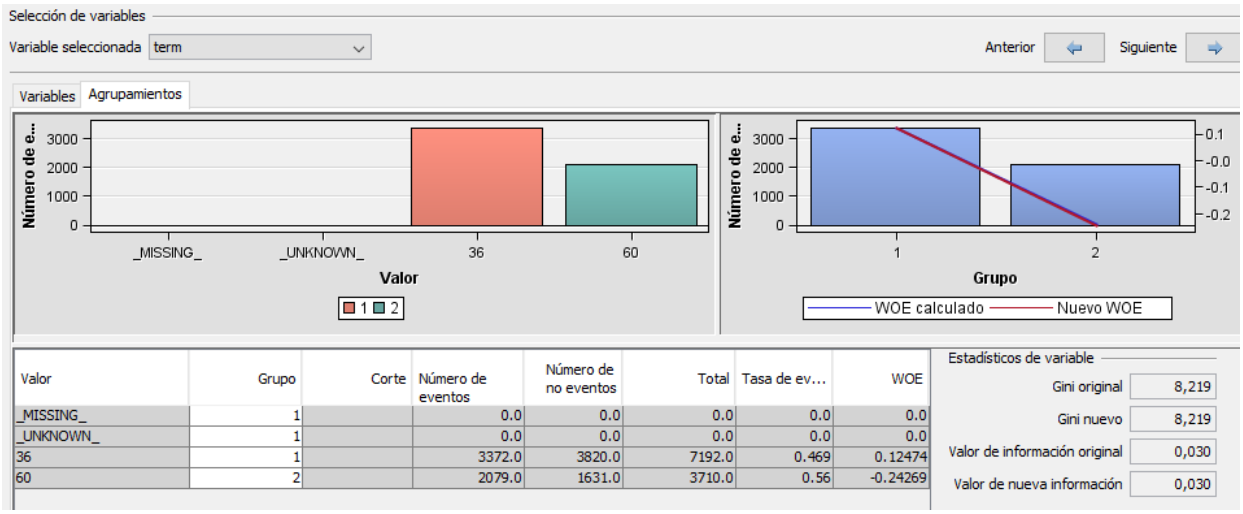


Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
UNKNOWN	1		0.0	0.0	0.0	0.0	0.0
ANY	1		3.0	0.0	3.0	1.0	-1.94591
MORTGAG	1		2101.0	2624.0	4725.0	0.445	0.22229
OWN	2		771.0	698.0	1469.0	0.525	-0.09947
RENT	3		2576.0	2129.0	4705.0	0.548	-0.19059

Estadísticos de variable

Gini original	10,067
Gini nuevo	10,067
Valor de información original	0,038
Valor de nueva información	0,038
Nivel de detalle	

Variable term

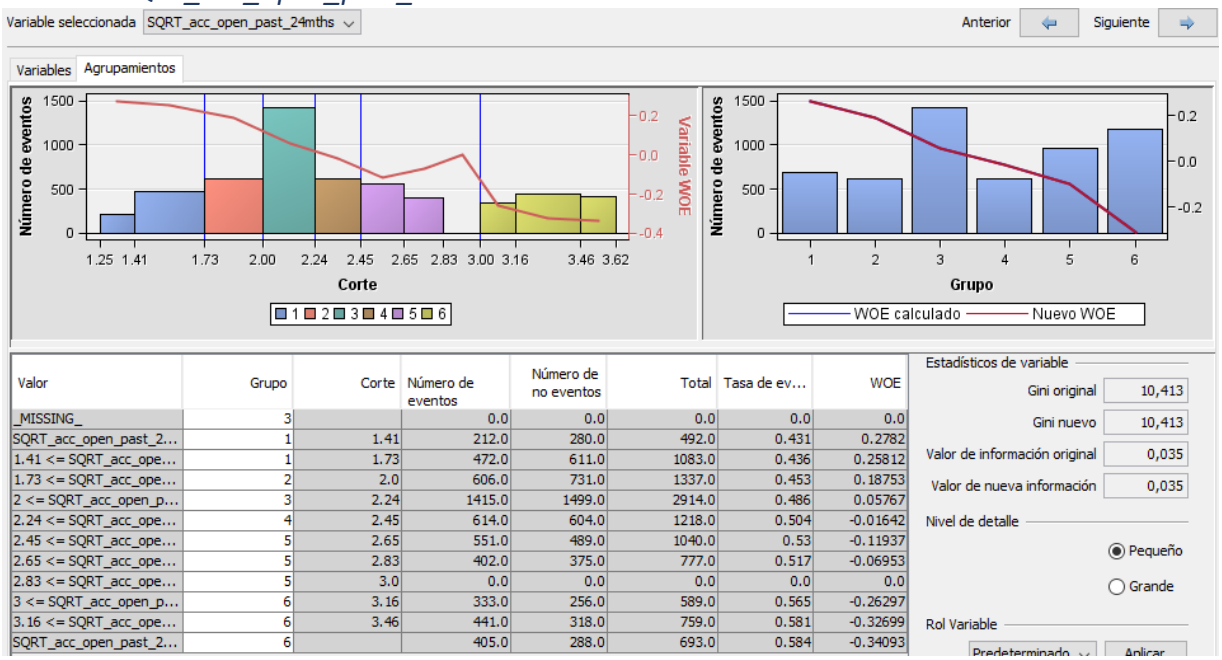


Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
UNKNOWN	1		0.0	0.0	0.0	0.0	0.0
36	1		3372.0	3820.0	7192.0	0.469	0.12474
60	2		2079.0	1631.0	3710.0	0.56	-0.24269

Estadísticos de variable

Gini original	8,219
Gini nuevo	8,219
Valor de información original	0,030
Valor de nueva información	0,030

Variable SQRT_acc_open_past_24mths



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	3		0.0	0.0	0.0	0.0	0.0
SQRT_acc_open_past_2...	1	1.41	212.0	280.0	492.0	0.431	0.2782
1.41 <= SQRT_acc_ope...	1	1.73	472.0	611.0	1083.0	0.436	0.25812
1.73 <= SQRT_acc_ope...	2	2.0	606.0	731.0	1337.0	0.453	0.18753
2 <= SQRT_acc_open_p...	3	2.24	1415.0	1499.0	2914.0	0.486	0.05767
2.24 <= SQRT_acc_ope...	4	2.45	614.0	604.0	1218.0	0.504	-0.01642
2.45 <= SQRT_acc_ope...	5	2.65	551.0	489.0	1040.0	0.53	-0.11937
2.65 <= SQRT_acc_ope...	5	2.83	402.0	375.0	777.0	0.517	-0.06953
2.83 <= SQRT_acc_ope...	5	3.0	0.0	0.0	0.0	0.0	0.0
3 <= SQRT_acc_open_p...	6	3.16	333.0	256.0	589.0	0.565	-0.26297
3.16 <= SQRT_acc_ope...	6	3.46	441.0	318.0	759.0	0.581	-0.32699
SQRT_acc_open_past_2...	6		405.0	288.0	693.0	0.584	-0.34093

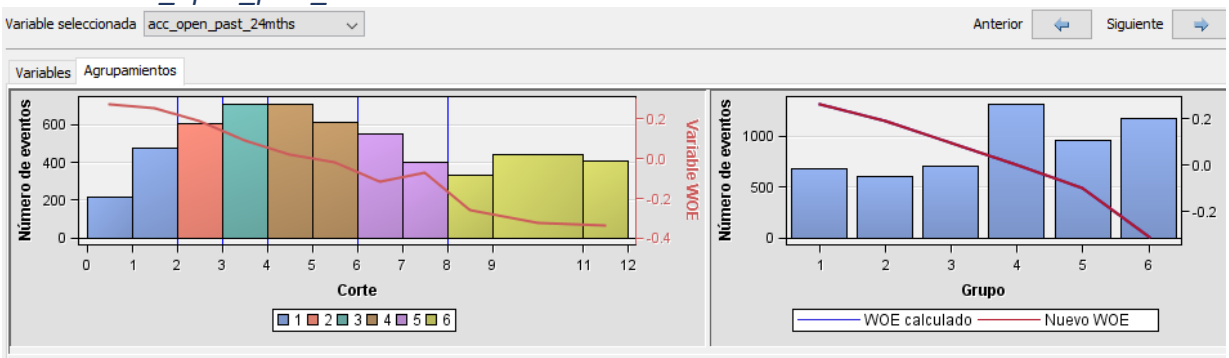
Estadísticos de variable

Gini original	10,413
Gini nuevo	10,413
Valor de información original	0,035
Valor de nueva información	0,035

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado / Aplicar

Variable acc_open_past_24mths



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
acc_open_past_24mths ...	1	1.0	212.0	280.0	492.0	0.431	0.2782
1 <= acc_open_past_24...	1	2.0	472.0	611.0	1083.0	0.436	0.25812
2 <= acc_open_past_24...	2	3.0	606.0	731.0	1337.0	0.453	0.18753
3 <= acc_open_past_24...	3	4.0	710.0	781.0	1491.0	0.476	0.09531
4 <= acc_open_past_24...	4	5.0	705.0	718.0	1423.0	0.495	0.01827
5 <= acc_open_past_24...	4	6.0	614.0	604.0	1218.0	0.504	-0.01642
6 <= acc_open_past_24...	5	7.0	551.0	489.0	1040.0	0.53	-0.11937
7 <= acc_open_past_24...	5	8.0	402.0	375.0	777.0	0.517	-0.06953
8 <= acc_open_past_24...	6	9.0	333.0	256.0	589.0	0.565	-0.26297
9 <= acc_open_past_24...	6	11.0	441.0	318.0	759.0	0.581	-0.32699
acc_open_past_24mths ...	6		405.0	288.0	693.0	0.584	-0.34093

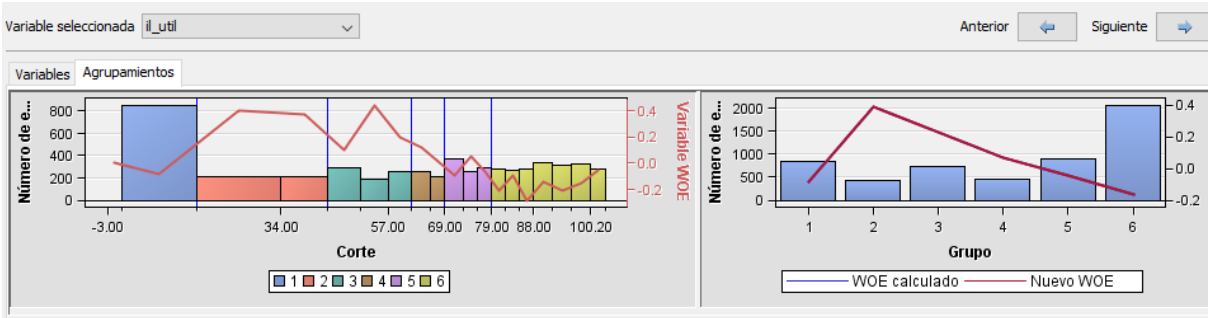
Estadísticos de variable

Gini original	10,500
Gini nuevo	10,500
Valor de información original	0,035
Valor de nueva información	0,035

Nivel de detalle: Pequeño / Grande

Rol Variable

Variable il_util



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	6		0.0	0.0	0.0	0.0	0.0
il_util < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= il_util < 16	1	16.0	852.0	783.0	1635.0	0.521	-0.08445
16 <= il_util < 34	2	34.0	210.0	314.0	524.0	0.401	0.40229
34 <= il_util < 44	2	44.0	206.0	299.0	505.0	0.408	0.37257
44 <= il_util < 51.07	3	51.07	289.0	318.0	607.0	0.476	0.09562
51.07 <= il_util < 57	3	57.0	188.0	293.0	481.0	0.391	0.44373
57 <= il_util < 62	3	62.0	259.0	314.0	573.0	0.452	0.19256
62 <= il_util < 66	4	66.0	258.0	290.0	548.0	0.471	0.11692
66 <= il_util < 69	4	69.0	212.0	215.0	427.0	0.496	0.01405
69 <= il_util < 73	5	73.0	365.0	332.0	697.0	0.524	-0.09476
73 <= il_util < 76	5	76.0	252.0	265.0	517.0	0.487	0.0503
76 <= il_util < 79	5	79.0	286.0	267.0	553.0	0.517	-0.06874
79 <= il_util < 82	6	82.0	281.0	227.0	508.0	0.553	-0.2134
82 <= il_util < 85	6	85.0	263.0	240.0	503.0	0.523	-0.09152
85 <= il_util < 88	6	88.0	283.0	213.0	496.0	0.571	-0.28415
88 <= il_util < 92	6	92.0	336.0	290.0	626.0	0.537	-0.14723
92 <= il_util < 96	6	96.0	307.0	248.0	555.0	0.553	-0.21342
96 <= il_util < 100.2	6	100.2	323.0	278.0	601.0	0.537	-0.15003
il_util >= 100.2	6		281.0	265.0	546.0	0.515	-0.05862

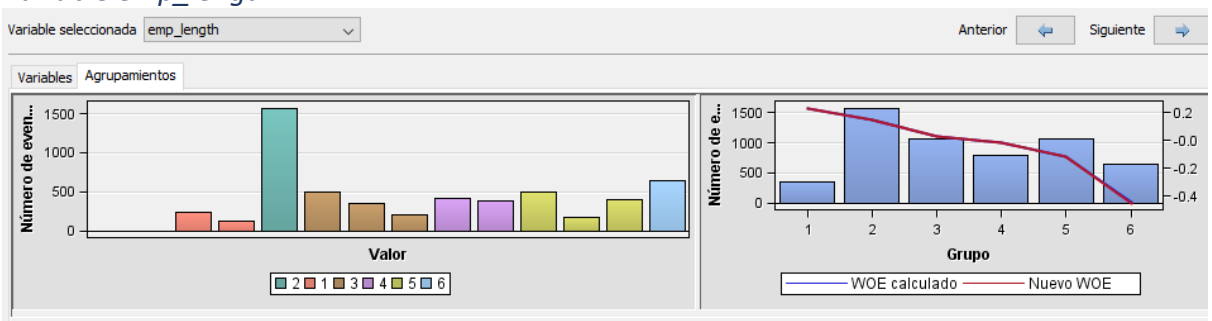
Estadísticos de variable

Gini original	9,614
Gini nuevo	9,614
Valor de información original	0,033
Valor de nueva información	0,033

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Variable emp_length



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	2		0.0	0.0	0.0	0.0	0.0
UNKNOWN	2		0.0	0.0	0.0	0.0	0.0
6 YEARS	1		227.0	293.0	520.0	0.437	0.25522
9 YEARS	1		117.0	141.0	258.0	0.453	0.18659
10+ YEARS	2		1569.0	1817.0	3386.0	0.463	0.14675
2 YEARS	3		500.0	507.0	1007.0	0.497	0.0139
5 YEARS	3		346.0	355.0	701.0	0.494	0.02568
7 YEARS	3		202.0	217.0	419.0	0.482	0.07163
1 YEAR	4		406.0	404.0	810.0	0.501	-0.00494
4 YEARS	4		376.0	364.0	740.0	0.508	-0.03244
3 YEARS	5		494.0	428.0	922.0	0.536	-0.14341
8 YEARS	5		173.0	160.0	333.0	0.52	-0.07812
< 1 YEAR	5		399.0	356.0	755.0	0.528	-0.11403
N/A	6		642.0	409.0	1051.0	0.611	-0.45087

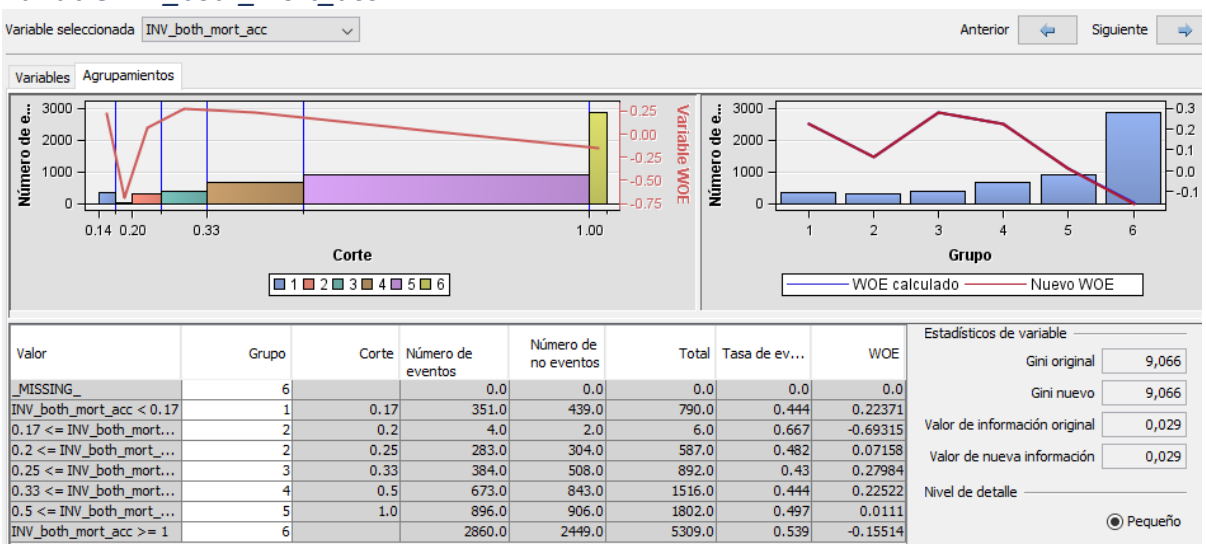
Estadísticos de variable

Gini original	9,376
Gini nuevo	9,376
Valor de información original	0,033
Valor de nueva información	0,033

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Variable INV_both_mort_acc



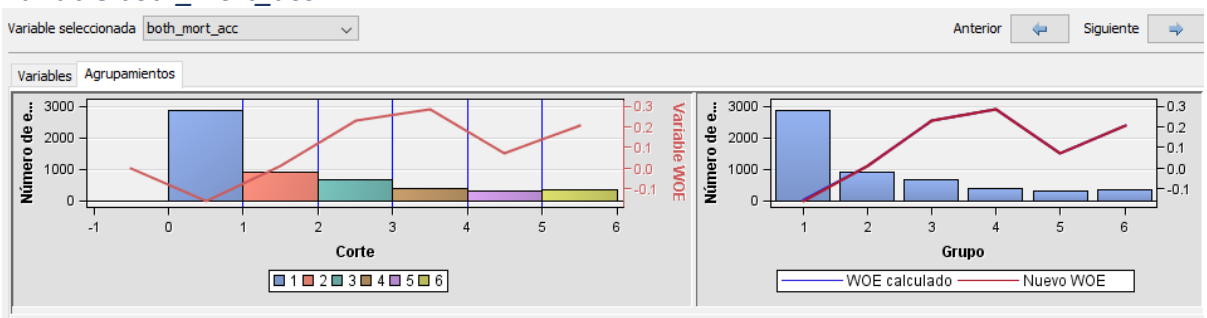
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	6		0.0	0.0	0.0	0.0	0.0
INV_both_mort_acc < 0.17	1	0.17	351.0	439.0	790.0	0.444	0.22371
0.17 <= INV_both_mort...	2	0.2	4.0	2.0	6.0	0.667	-0.69315
0.2 <= INV_both_mort_...	2	0.25	283.0	304.0	587.0	0.482	0.07158
0.25 <= INV_both_mort...	3	0.33	384.0	508.0	892.0	0.43	0.27984
0.33 <= INV_both_mort...	4	0.5	673.0	843.0	1516.0	0.444	0.22522
0.5 <= INV_both_mort_...	5	1.0	896.0	906.0	1802.0	0.497	0.0111
INV_both_mort_acc >= 1	6		2860.0	2449.0	5309.0	0.539	-0.15514

Estadísticos de variable

Gini original	9,066
Gini nuevo	9,066
Valor de información original	0,029
Valor de nueva información	0,029

Nivel de detalle: Pequeño

Variable both_mort_acc



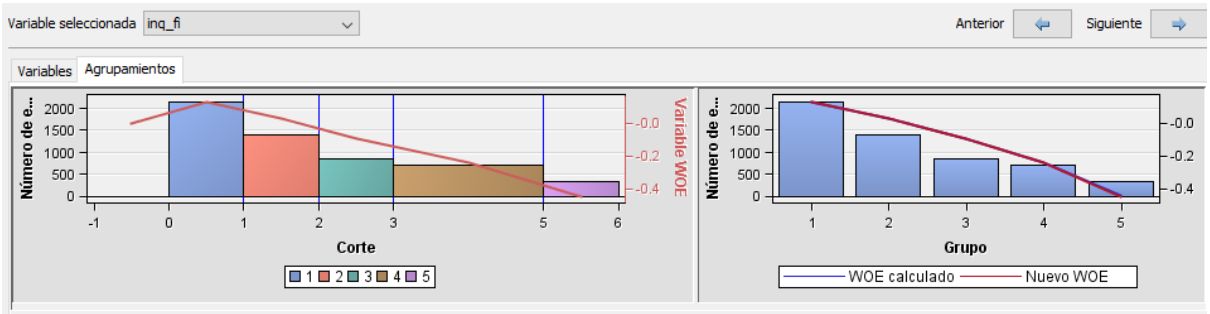
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
both_mort_acc < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= both_mort_acc < 1	1	1.0	2860.0	2449.0	5309.0	0.539	-0.15514
1 <= both_mort_acc < 2	2	2.0	898.0	907.0	1805.0	0.498	0.00997
2 <= both_mort_acc < 3	3	3.0	671.0	844.0	1515.0	0.443	0.22938
3 <= both_mort_acc < 4	4	4.0	384.0	510.0	894.0	0.43	0.28377
4 <= both_mort_acc < 5	5	5.0	289.0	311.0	600.0	0.482	0.07337
both_mort_acc >= 5	6		349.0	430.0	779.0	0.448	0.20871

Estadísticos de variable

Gini original	9,094
Gini nuevo	9,094
Valor de información original	0,029
Valor de nueva información	0,029

Nivel de detalle: Pequeño

Variable inq_fi



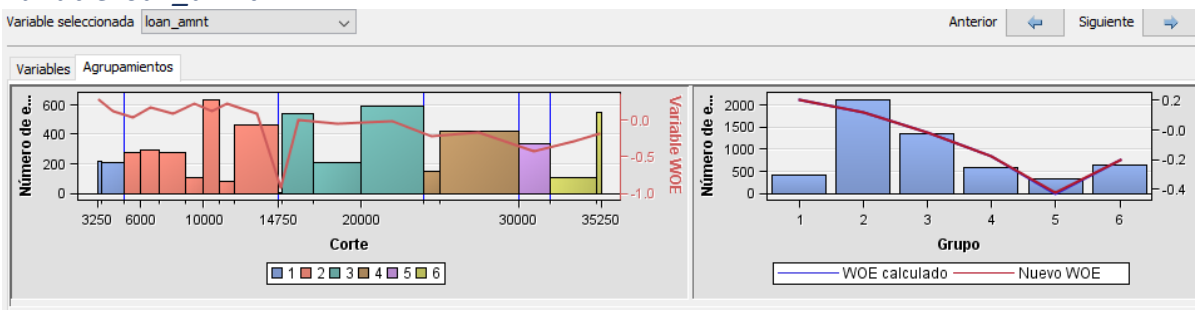
Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	1		0.0	0.0	0.0	0.0	0.0
inq_fi < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= inq_fi < 1	1	1.0	2162.0	2467.0	4629.0	0.467	0.13197
1 <= inq_fi < 2	2	2.0	1401.0	1444.0	2845.0	0.492	0.03023
2 <= inq_fi < 3	3	3.0	843.0	768.0	1611.0	0.523	-0.09318
3 <= inq_fi < 5	4	5.0	711.0	558.0	1269.0	0.56	-0.24231
inq_fi >= 5	5		334.0	214.0	548.0	0.609	-0.44516

Estadísticos de variable

Gini original	8,135
Gini nuevo	8,135
Valor de información original	0,026
Valor de nueva información	0,026

Nivel de detalle

Variable loan_amnt



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	2		0.0	0.0	0.0	0.0	0.0
loan_amnt < 3500	1	3500.0	213.0	284.0	497.0	0.429	0.28768
3500 <= loan_amnt < 5...	1	5000.0	210.0	236.0	446.0	0.471	0.11672
5000 <= loan_amnt < 6...	2	6000.0	274.0	283.0	557.0	0.492	0.03232
6000 <= loan_amnt < 7...	2	7200.0	294.0	349.0	643.0	0.457	0.17149
7200 <= loan_amnt < 8...	2	8900.0	279.0	303.0	582.0	0.479	0.08252
8900 <= loan_amnt < 1...	2	10000.0	105.0	132.0	237.0	0.443	0.22884
10000 <= loan_amnt < ...	2	11000.0	636.0	715.0	1351.0	0.471	0.11708
11000 <= loan_amnt < ...	2	12000.0	80.0	100.0	180.0	0.444	0.22314
12000 <= loan_amnt < ...	2	14750.0	458.0	500.0	958.0	0.478	0.08774
14750 <= loan_amnt < ...	3	15000.0	5.0	2.0	7.0	0.714	-0.91629
15000 <= loan_amnt < ...	3	17000.0	541.0	542.0	1083.0	0.5	0.00185
17000 <= loan_amnt < ...	3	20000.0	209.0	199.0	408.0	0.512	-0.04903
20000 <= loan_amnt < ...	3	24000.0	593.0	581.0	1174.0	0.505	-0.02044
24000 <= loan_amnt < ...	4	25000.0	150.0	121.0	271.0	0.554	-0.21484
25000 <= loan_amnt < ...	4	30000.0	418.0	355.0	773.0	0.541	-0.16336
30000 <= loan_amnt < ...	5	32000.0	339.0	221.0	560.0	0.605	-0.42784
32000 <= loan_amnt < ...	6	35000.0	102.0	77.0	179.0	0.57	-0.28117
loan_amnt >= 35000	6		545.0	451.0	996.0	0.547	-0.18932

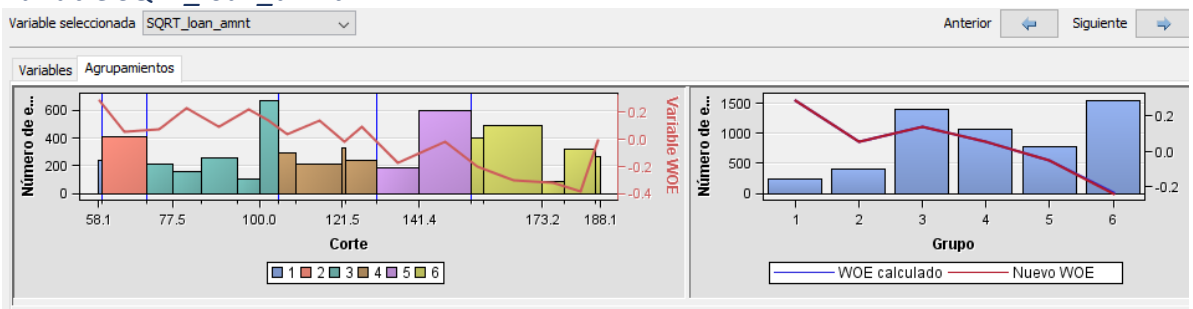
Estadísticos de variable

Gini original	8,441
Gini nuevo	8,441
Valor de información original	0,026
Valor de nueva información	0,026

Nivel de detalle: Pequeño / Grande

Rol Variable: Predeterminado — Aplicar

Variable SQRT_loan_amnt



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	3		0.0	0.0	0.0	0.0	0.0
SQRT_loan_amnt < 59.17	1	59.17	240.0	321.0	561.0	0.428	0.2908
59.17 <= SQRT_loan_a...	2	70.72	407.0	430.0	837.0	0.486	0.05497
70.72 <= SQRT_loan_a...	3	77.47	212.0	228.0	440.0	0.482	0.07276
77.47 <= SQRT_loan_a...	3	84.86	159.0	200.0	359.0	0.443	0.22941
84.86 <= SQRT_loan_a...	3	94.35	254.0	279.0	533.0	0.477	0.09388
94.35 <= SQRT_loan_a...	3	100.0	103.0	129.0	232.0	0.444	0.22508
100 <= SQRT_loan_amn...	3	104.89	672.0	772.0	1444.0	0.465	0.13873
104.89 <= SQRT_loan_...	4	109.55	293.0	303.0	596.0	0.492	0.03356
109.55 <= SQRT_loan_...	4	121.45	209.0	240.0	449.0	0.465	0.1383
121.45 <= SQRT_loan_...	4	122.48	331.0	324.0	655.0	0.505	-0.02137
122.48 <= SQRT_loan_...	4	130.39	239.0	263.0	502.0	0.476	0.09569
130.39 <= SQRT_loan_...	5	141.42	185.0	156.0	341.0	0.543	-0.1705
141.42 <= SQRT_loan_...	5	154.92	593.0	581.0	1174.0	0.505	-0.02044
154.92 <= SQRT_loan_...	6	158.12	396.0	323.0	719.0	0.551	-0.20376
158.12 <= SQRT_loan_...	6	173.21	486.0	358.0	844.0	0.576	-0.30568
173.21 <= SQRT_loan_...	6	178.89	88.0	64.0	152.0	0.579	-0.31845
178.89 <= SQRT_loan_...	6	187.09	318.0	216.0	534.0	0.596	-0.38677
SQRT_loan_amnt >= 18...	6		266.0	264.0	530.0	0.502	-0.00755

Estadísticos de variable

Gini original	8,643
Gini nuevo	8,643
Valor de información original	0,025
Valor de nueva información	0,025

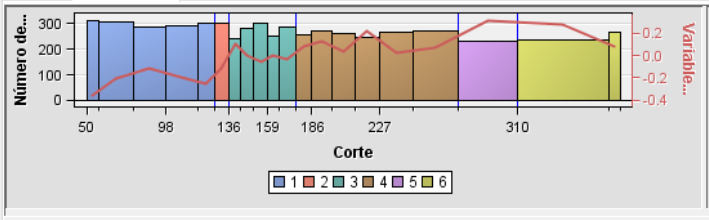
Nivel de detalle: Pequeño / Grande

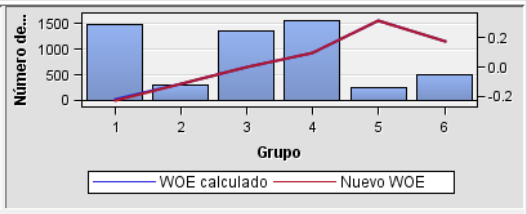
Rol Variable: Predeterminado — Aplicar

Variable mths_earliest_cr_line

Variable seleccionada: mths_earliest_cr_line — Anterior — Siguiente

Variables | Agrupamientos





Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de e...	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
mths_earliest_cr_line < 57	1	57.0	314.0	220.0	534.0	0.588	-0.35577
57 <= mths_earliest_cr...	1	78.0	304.0	246.0	550.0	0.553	-0.2117
78 <= mths_earliest_cr...	1	98.0	285.0	255.0	540.0	0.528	-0.11123
98 <= mths_earliest_cr...	1	117.0	289.0	238.0	527.0	0.548	-0.19416
117 <= mths_earliest_c...	1	127.0	301.0	233.0	534.0	0.564	-0.25607
127 <= mths_earliest_c...	2	136.0	300.0	267.0	567.0	0.529	-0.11653
136 <= mths_earliest_c...	3	143.0	239.0	266.0	505.0	0.473	0.10703
143 <= mths_earliest_c...	3	151.0	279.0	278.0	557.0	0.501	-0.00359
151 <= mths_earliest_c...	3	159.0	303.0	286.0	589.0	0.514	-0.05774
159 <= mths_earliest_c...	3	166.0	249.0	250.0	499.0	0.499	0.00401
166 <= mths_earliest_c...	3	176.0	287.0	277.0	564.0	0.509	-0.03546
176 <= mths_earliest_c...	4	186.0	256.0	278.0	534.0	0.479	0.08244
186 <= mths_earliest_c...	4	198.0	272.0	310.0	582.0	0.467	0.13077
198 <= mths_earliest_c...	4	212.0	259.0	267.0	526.0	0.492	0.03042
212 <= mths_earliest_c...	4	227.0	246.0	306.0	552.0	0.446	0.21825
227 <= mths_earliest_c...	4	247.0	267.0	272.0	539.0	0.495	0.01855
247 <= mths_earliest_c...	4	274.0	271.0	290.0	561.0	0.483	0.06776
274 <= mths_earliest_c...	5	310.0	231.0	317.0	548.0	0.422	0.31648
310 <= mths_earliest_c...	6	365.0	234.0	307.0	541.0	0.433	0.27153
mths_earliest_cr_line >...	6		265.0	288.0	553.0	0.479	0.08323

Estadísticos de variable

Gini original	8,449
Gini nuevo	8,449
Valor de información original	0,024
Valor de nueva información	0,024

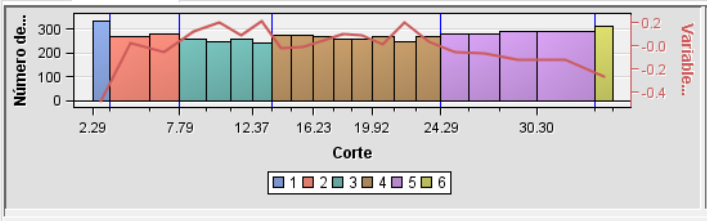
Nivel de detalle: (•) Pequeño ( ) Grande

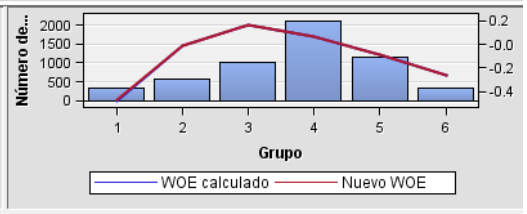
Rol Variable: Predeterminado — Aplicar

Variable dti_total

Variable seleccionada: dti_total — Anterior — Siguiente

Variables | Agrupamientos





Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de e...	WOE
MISSING	4		0.0	0.0	0.0	0.0	0.0
dti_total < 3.42	1	3.42	336.0	209.0	545.0	0.617	-0.47478
3.42 <= dti_total < 5.87	2	5.87	268.0	276.0	544.0	0.493	0.02941
5.87 <= dti_total < 7.79	2	7.79	281.0	265.0	546.0	0.515	-0.05862
7.79 <= dti_total < 9.51	3	9.51	255.0	290.0	545.0	0.468	0.12862
9.51 <= dti_total < 10.99	3	10.99	244.0	301.0	545.0	0.448	0.20994
10.99 <= dti_total < 12...	3	12.37	260.0	284.0	544.0	0.478	0.08829
12.37 <= dti_total < 13...	3	13.59	242.0	302.0	544.0	0.445	0.22149
13.59 <= dti_total < 14...	4	14.83	275.0	269.0	544.0	0.506	-0.02206
14.83 <= dti_total < 16...	4	16.23	274.0	271.0	545.0	0.503	-0.01101
16.23 <= dti_total < 17...	4	17.54	267.0	281.0	548.0	0.487	0.05111
17.54 <= dti_total < 18...	4	18.67	258.0	287.0	545.0	0.473	0.10652
18.67 <= dti_total < 19...	4	19.92	258.0	283.0	541.0	0.477	0.09249
19.92 <= dti_total < 21...	4	21.29	270.0	275.0	545.0	0.495	0.01835
21.29 <= dti_total < 22...	4	22.69	244.0	301.0	545.0	0.448	0.20994
22.69 <= dti_total < 24...	4	24.29	270.0	279.0	549.0	0.492	0.03279
24.29 <= dti_total < 26...	5	26.05	280.0	265.0	545.0	0.514	-0.05506
26.05 <= dti_total < 27...	5	27.96	281.0	264.0	545.0	0.516	-0.06241
27.96 <= dti_total < 30.3	5	30.3	290.0	256.0	546.0	0.531	-0.1247
30.3 <= dti_total < 33.96	5	33.96	289.0	256.0	545.0	0.53	-0.12125
dti_total >= 33.96	6		309.0	237.0	546.0	0.566	-0.26528

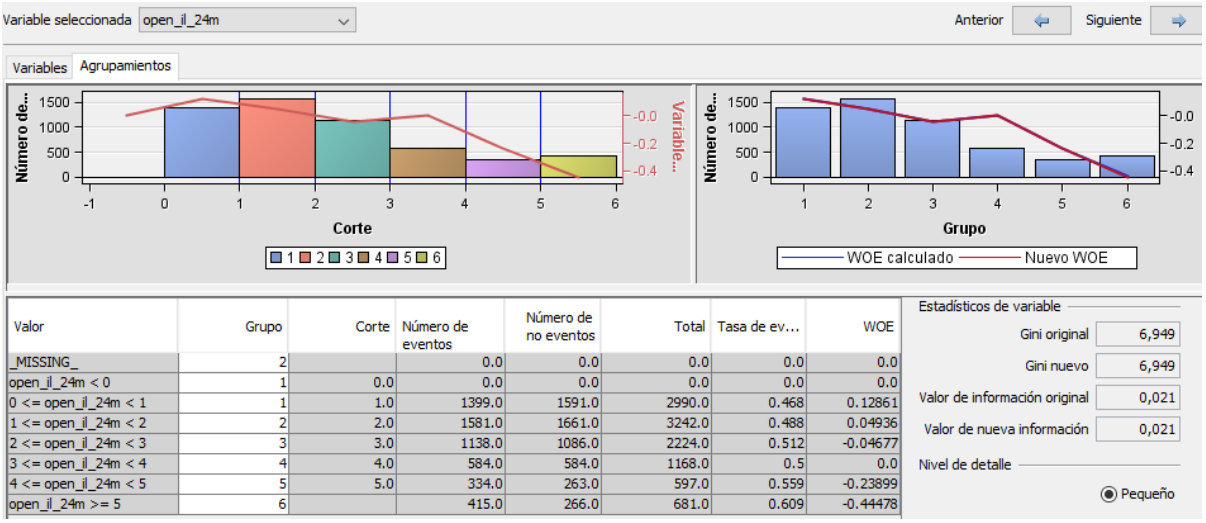
Estadísticos de variable

Gini original	7,531
Gini nuevo	7,531
Valor de información original	0,023
Valor de nueva información	0,023

Nivel de detalle: (•) Pequeño ( ) Grande

Rol Variable: Predeterminado — Aplicar

Variable open_il_24m



Valor	Grupo	Corte	Número de eventos	Número de no eventos	Total	Tasa de ev...	WOE
MISSING	2		0.0	0.0	0.0	0.0	0.0
open_il_24m < 0	1	0.0	0.0	0.0	0.0	0.0	0.0
0 <= open_il_24m < 1	1	1.0	1399.0	1591.0	2990.0	0.468	0.12861
1 <= open_il_24m < 2	2	2.0	1581.0	1661.0	3242.0	0.488	0.04936
2 <= open_il_24m < 3	3	3.0	1138.0	1086.0	2224.0	0.512	-0.04677
3 <= open_il_24m < 4	4	4.0	584.0	584.0	1168.0	0.5	0.0
4 <= open_il_24m < 5	5	5.0	334.0	263.0	597.0	0.559	-0.23899
open_il_24m >= 5	6		415.0	266.0	681.0	0.609	-0.44478

Estadísticos de variable

Gini original	6,949
Gini nuevo	6,949
Valor de información original	0,021
Valor de nueva información	0,021

Nivel de detalle: Pequeño

Anexo IV - Configuración de los modelos para el set sin tramificar

MEJORES MODELOS SET SIN TRAMIFICAR				
Red Neuronal = avnnet_st				
	Nodos	**Tasa Aprendizaje**	**Iteraciones**	
	11	0.1	100	
Random Forest= rf_st				
Árboles	**Tamaño Muestra**	**Obs. Mín. Nodo**	**Profundidad Max.**	**Variables Sortear**
500	TOTAL	50	10	8
Gradient Boosting paquete gbm = gbm_st				
Árboles	**Tasa Aprendizaje**	**Obs. Mín. Nodo**	**Profundidad Max.**	
2000	0.05	50	10	
Gradient Boosting paquete XGBoost = xgbm_st				
Árboles	**Tasa Aprendizaje**	**Obs. Mín. Nodo**	**Profundidad Max.**	
500	0.03	50	10	
gamma=0; colsample_bytree=0.8, subsample=1, alpha=0.1, lambda=0.5, l_bias=0				
SVMPolinómico= SVMPoly_st				
	Inverso Margen Error	**Grado del Polinomio**	**Escala**	
	0.01	2	2	
SVM Polinomial = SVMPoly_st				
	Inv. Margen Error	**Sigma**		
	5	0.01		

Anexo V- Acceso a los archivos de código utilizados

Con el siguiente link se puede acceder a la carpeta compartida en Google Drive donde se encuentra el desarrollo del trabajo en ambos SAS y R.

https://bit.ly/2Fv2Ngd